ИНСТИТУТ БИОИНФОРМАТИКИ

СБОРНИК ТЕЗИСОВ 2020/21

ПОЛИТЕХ-ПРЕСС
Санкт-Петербургский
политехнический университет
Петра Великого

Санкт-Петербург
2021

УДК 004.94,575.1,575.8,577.2,578.5,579.25
И71

Институт биоинформатики. Сборник тезисов 2020/21. – СПб. : ПОЛИТЕХ-ПРЕСС, 2021. – 87 с.

Результаты НИР 2020/21 учебного года.
Летняя школа по биоинформатике 2021. Тезисы докладов.

© Институт биоинформатики, 2021
© Санкт-Петербургский политехнический университет Петра Великого, 2021

ISBN 978-5-7422-7387-5

BIOINFORMATICS INSTITUTE
RESEARCH PROJECTS 2020/21

Saint-Petersburg
2021

Bioinformatics Institute research projects 2020/21. – Saint-Petersburg. : POLYTECH-PRESS, 2021. – 87 p.

Bioinformatics Institute 2020/21. Project abstracts.
Bioinformatics Summer School 2021. Abstracts.

© Bioinformatics Institute, 2021
© Peter the Great St. Petersburg State
Polytechnic University, 2021

ISBN 978-5-7422-7387-5

Table of contents

FALL 2020 8

Modeling mutations in SARS-CoV-2 Spike protein 9

Assembly and analysis of organelle genomes of *Iris* species and *Picea abies* ... 10

Esophageal cancer incidence and mortality trends in Russia 1993-2018 11

Comparing bioinformatic pipelines for microbiome data analysis 12

Analysis of differential expression of cryptic vole species and voles from various habitats 13

Uncovering molecular characteristics of cellular senescence process specific to human mesenchymal stem cells 14

Simulation of DNA-probe and DNA-target hybridization in Tyramide-FISH method 15

Numerical Monte-Carlo solution to the system of stochastic differential equations of population dynamics 16

GWAS analysis of region of genome associated with severe Covid-19 17

Prediction protein sequences using machine learning approach 18

Orientations of contigs using Hi-C data 19

Annotation of metagenomes of the microbial cellulolytic communities 20

Scaffolding based on Hi-C reads and Neural Networks 21

Analysis of batch effect 22

Agent-based modeling of the spread of antibiotic resistance in a bacterial population 23

Improve casual gene prioritization by patient's phenotype. Identification and parsing of relevant Mendelian disease articles 24

Study and improvement of label propagation binning algorithm 25

Analysis of differential gene expression of non-model hydroid polyp *Dynamena pumila* 26

Study of transcriptome during intense exercises at high altitude 27

Comprehensive analysis of shared genetic architecture between neurological and psychiatric disorders 28

Assembly and analysis of *Preeria caryophylla* genome 30

Improve casual gene prioritization by patient's phenotype 31

Bacterial genome assembly and decontamination 33

Search for molecular markers - predictors of a positive response to immunotherapeutic treatment according to single-cell RNA sequencing 34

Detecting novel molecular events in proteomics data for genetic diagnostics ... 35
Hidden biodiversity: search for uncultured protists in metagenomes ... 36
Statistical estimation of von Willebrand factor exocytosis from human endothelial cells ... 37
Lipophilicity prediction with graph convolutions and molecular substructures representation ... 39
Epidemic simulation of COVID-19 ... 41
Analysis of metagenomic samples from black taiga and regional soil ... 42
Search for frame shift signals in bacterial genomes ... 43
Comparison of methods for analysis of differential gene expression ... 44
In search of PET-degrading enzymes: metagenomes reveal hidden homologues ... 46
SPRING 2021 ... 48
Shared biologic architecture between Parkinson’s disease and inflammatory disorders ... 49
Identification and comparison of somatic antigen structures of symbiotic and pathogenic bacteria from *Morganellaceae* family ... 50
Comparative analysis of methods for batch correction in proteomics ... 51
Studying role of rare variants on whole genome sequencing data of Parkinson’s disease (PD) patients ... 52
Identification and analysis of SARS-CoV-2-specific T-cell receptors ... 53
Adding context to nonsense: analysis of sequence properties at pLoF variant sites ... 54
Ordering of contigs using Hi-C data ... 55
Search for convergent adaptive substitutions in alpine rodents ... 56
Evaluation of DNA methylation episignatures ... 58
Age-related changes in methylation patterns at sites of active human DNA replication ... 59
Analysis of transcriptome signatures of patients with heart failure and with myopathy ... 61
Analysis of evolutionary conservation of polyproline motifs ... 63
Influence of microbiota on the effectiveness of immunotherapeutic drugs in the treatment of metastatic solid tumors ... 64
Developing best practices for single-cell analysis: data integration ... 65

Fine-mapping of causal variants in the *SCARB2* and *CTSB* genes in Parkinson's disease..66

Preoperative assessment of glioblastoma resectability and development of a phenotypic rating scale..68

Comparative analysis of genomic assemblies, repertoire of repeats, and protein-coding genes of palm trees *Elaeis guineensis* (*pisifera*) and *Elaeis oleifera*......69

Bioinformatics analysis of E3 ubiquitin ligase family......................................70

Inference of inbreeding coefficients in GADMA 71

Search for the insertion site of human betaherpesvirus 6A into the human chromosome 72

Building gene regulatory networks with structural properties...........................73

Haplotype variant calling using hypergraphs approach....................................74

Transcriptomics and small RNAome responses to the infection by phytopathogenic fungi: an analysis in the wild model legume *Medicago truncatula* infected by *Verticillium alfalfae* 75

BIOINFORMATICS SUMMER SCHOOL 2021.......................................76

Филогения *Lumbrineridae* (*Annelida*) на основе молекулярных данных......77

Using protein-protein interaction and gene networks to improve Connectivity Map..78

Веб-сервис WebMCOT для определения совместно встречаемых ДНК мотивов в данных CHiP-Seq...79

Risk of mitochondrial deletions is affected by the global secondary structure of the mitochondrial genome...80

HLA variants differ in binding preferences of self-peptides from proteins with specific molecular functions...82

TCRen: a knowledge-based pairwise potential that accurately predicts TCR-peptide recognition...84

Триплетный состав общих генов митохондрий и хлоропластов растений выявляет их дифференциацию..85

Филогенетическая классификация и биосинтетический потенциал штамма *Rhodococcus* S11...86

FALL 2020

УДК 577.2, 578.5

Modeling mutations in SARS-CoV-2 Spike protein

O. Vavulov[1,2], E. Ivanova[2,3], A. Shemyakina[2,4], K. Varchenko[2,5],
M. Akimenkova[2,6], L. Danilov[2,7], A. Zolotarev[2]

[1] Sberbank PJSC, Vavilova St. 19, Moscow, 117997, Russia
[2] Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[3] "GenBit" LLC, Varshavskoye Highway 28A, 117105, Moscow, Russia
[4] Scientific Center "Kurchatov Institute" Research Institute for Genetics and Selection of Industrial Microorganisms, 1-st Dorozhniy pr. 1, 117545, Moscow, Russia
[5] Covance, Serdobolskaya St. 64, 197342, St. Petersburg, Russia
[6] Moscow Institute of Physics and Technology, Institutskiy pk. 9, 141701, Dolgoprudniy, Russia
[7] St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia

SARS-CoV2 virus caused a pandemic with more than a million deaths, and the number of deaths is constantly growing. The studies on virus receptor binding domain (RBD) revealed its increased affinity to angiotensin converting enzyme 2 (ACE2) compared to SARS-CoV that caused the 2002-2004 SARS outbreak. Thus, we attempted to simulate further evolution of RBD in this direction and suggest the class of drugs that could inhibit mutant RBD (RBD-mut).

We analysed the RBD-ACE2 interface in the PyMOL and identified key amino acid residues of their interaction. Using the FoldX based pipeline we went through all possible missense mutations in these codons and selected combinations of them that: a) preserve the stability of RBD; b) increase the stability of the RBD/ACE2 complex. Two RBD-muts with high affinity were chosen for molecular docking - RBD-mut-1 (Y489F;Q498L) and RBD-mut-2 (Y453F;N487K;Y489F;Q498L). Our results show that one of the ways of the virus evolution can possibly be associated with an increase of RBD hydrophobicity.

Using AutoDock based pipeline, FDA-approved molecules were docked against the selected RBD-muts, and the resulting complexes were ranked by interaction energy. Possible intermolecular interactions in the first 100 RBD-mut/ligand complexes for each RBD-mut were manually analyzed in PyMOL. The molecules that could reliably bind the interface region by the formation of polar contacts and hydrophobic interactions were suggested as potentially effective competitive inhibitors of each RBD-mut. For RBD-mut-1 molecules with heterocyclic core and polar groups on the sides were chosen as possible inhibitors. For RBD-mut-2 molecules with aromatic parts instead of heterocycles were suggested.

УДК 575.8, 577.2

Assembly and analysis of organelle genomes of *Iris* species and *Picea abies*

A. Andreev[1,2,3], P. Zhurbenko[1,4], L. Danilov[1,5], M. Raiko[1,5]

[1] Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[2] St. Petersburg Forestry Research Institute, Institutskiy pr. 21, 194021, St. Petersburg, Russia
[3] St. Petersburg State Pediatric Medical University, Litovskaya st. 2, 194100, St. Petersburg, Russia
[4] Komarov Botanical Institute of the Russian Academy of Sciences, Professor Popov st. 2, 197376, St. Petersburg, Russia
[5] St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia

The aim of our work was to assemble and analyze chloroplast genomes of four *Iris* species (*I. lycotis, I. munzii, I. filifolia* and *I. pamphilica)* and mitochondrial genome of *Picea abies*. After assembling the genomes we constructed a phylogenetic tree of irises and searched for NUMTs in *P. abies* genome. Paired-end Illumina reads of Iris species were trimmed and filtered using Trimmomatic and then assembled using SPAdes. The assemble quality control was performed by aligning the scaffolds on the reference genome of *I. gatesii* using Quast and BWA. Well-assembled scaffolds along with homologous sequences of Iris species available in GenBank were used to construct phylogeny. The final 70 kbp alignment included 10 species belonging to four Iris subgenera. The phylogeny was constructed in the IQ-tree. Our data confirmed the phylogeny of the studied species, obtained earlier by five chloroplast genes [1]. To study the *P. abies* genome, we used data from Nystedt *et. al* 2013 [2]. Based on earlier gel electrophoresis analysis, it was shown that the *NAD1* gene is presented in the genome in multiple copies and can be a NUMT. To test this hypothesis, we selected scaffolds containing this gene using BLAST and QUAST. Several copies of this gene were found, confirming the PCR data, but they were located on scaffolds less than 1kbp in length, which made it impossible to conclude whether these belongs to the mitochondrial or nuclear genome.

References

1. Wilson C. A. Subgeneric classification in Iris re-examined using chloroplast sequence data. Taxon, 2011, 60(1), 27–35.
2. Nystedt B., et al. The Norway spruce genome sequence and conifer genome evolution. Nature, 2013, 497(7451), 579–584.

УДК 314.4

Esophageal cancer incidence and mortality trends in Russia 1993-2018

D. Andreeva[1,2], L. Danilov[1,2]

[1] St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia
[2] Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia

Esophageal cancer (EC) is the eighth most common cancer worldwide and the sixth leading cause of cancer related mortality [1]. The aim of this study is to analyze changes in EC incidence and mortality trends using data from the Russian State Cancer Registry.

Age-standardized rates (ASR) of cancer incidence and mortality per 100,000 person-years were calculated using the Segi-Doll world standard population [2]. Time-series analysis provides determination of the nature of the series and makes predictions of future values of the time series based on the present and past values [3].

During the study period, an average 8027 cases of EC were registered in Russia annually, with 185,822 deaths from the disease, rates for men are about 6 times higher than those for women. A steady decrease from 1993 to 2006 (women) or 2008 (men) and following increase were observed.

An increasing risk was observed with age for EC incidence and mortality, age effects plateaued at age 80-84 and above for women and 70-74 and then decreased for men. Due to the increase in incidence over the past 10 years, predictions also showed an increase in cases until 2028.

The increasing risk of EC can partly be explained by changes in the current Siewert classification and continuing trends of population aging and growth [4]. Increasing numbers of EC incident cases should be taken into account in planning healthcare resources at the national level.

References

1. Ferlay J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer, 2015, 136(5), E359–E386.
2. Doll R., Cook P. Summarizing indices for comparison of cancer incidence data. International journal of cancer, 1967, 2(3), 269–279.
3. Brockwell P. J. Time series analysis. Wiley StatsRef: Statistics Reference Online, 2014.
4. Rice T. W., Blackstone E. H., Rusch V. W. of the AJCC Cancer Staging Manual: esophagus and esophagogastric junction. Annals of surgical oncology, 2010, 17(7), 1721–1724.

УДК 579.25

Comparing bioinformatic pipelines for microbiome data analysis

A. Churkina[1], L. Danilov[2,3], M. Rayko[2,3]

[1] Almazov National Medical Research Centre, Akkuratova st. 2, 197341, St. Petersburg, Russia
[2] St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia
[3] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia

Metagenomics is a rapidly developing area of genomics, the task of which is the phylogenetic characterization of microbial communities, including strains that are uncultured. Metagenomic analysis of next generation sequencing (NGS) data of the 16S rRNA marker gene or whole metagenome sequencing is widely used in environmental and biomedical research.

To date, a large number of methods have been developed to analyze metagenomic data. In order to determine the most optimal method for analyzing metagenomic data, this study compared the 4 most popular bioinformatics pipelines: USEARCH-UPARSE(OTUs), USEARCH-UNOISE3(ASVs), Qiime2-Deblur(ASVs), and DADA2(ASVs). Taxonomic composition, as well as α- and β-diversity were determined with each pipeline using 77 Illumina paired-end read datasets (2x250) from obese and healthy controls.

Despite the good reproducibility of all methods in determining taxonomy at the levels of Phylums and Classes, as well as the beta diversity(Bray-Curtis PCoA), in the OTU method, the number of taxonomic units is less than in the ASV methods at all phylogenetic levels. The Usearch-Unoise3 pipeline contained a large number of "unrecognized" taxa, while Qiime2-Deblur had the lowest number of such taxa. In the DADA2 pipeline the maximum number of taxa at the levels of Type, Class, Order was determined, which was also confirmed by the presence of a significant ancestor of alpha-diversity indices, obtained as a result of the analysis using this pipeline.

More accurate and specific definition of taxonomy by ASV methods suggest that the ASV methods (such as DADA2 or Qiime2-Deblur) are the most preferred in the analysis of metagenomic data.

УДК 575.8, 577.2

Analysis of differential expression of cryptic vole species and voles from various habitats

K.V. Danko[1,2], S.A. Ilyutkin[1,2], K.S. Soghomonyan[1,2], A.V. Sidorin[1,2],
T.V. Petrova[3], S.Y. Bodrov[3], O.V. Bondareva[2,3]

[1] St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia
[2] Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[3] Department of Molecular Systematics, Laboratory of Theriology, Zoological Institute, Russian Academy of Sciences, Universitetskaya emb. 1, 199034, St. Petersburg, Russia

Habitat adaptation is one of the key factors in evolutionary success and species radiation. Voles inhabit different biotopes worldwide, and closely related species live in forests and in rocky mountains; moreover, cryptic species live in the same ecological niches. Vole adaptation to different habitats may occur due to changes in gene expression., though differential expression of genes between voles from different habitats and cryptic species from the same habitat has not been previously assessed.

We used differential expression analysis for close relative species, inhabiting forests (*Clethrionomys glareolus)*, forest mountains (*Chionomys nivalis*), rocky mountains (*Alticola lemminus)*, and two cryptic species *Lasiopodomys raddei* and *Lasiopodomys gregalis* from steppe. We found that voles from mountain habitats have higher levels of expressed genes involved in regulation of lipids metabolism, fibrinolysis and *Red/Ox* processes in comparison with *C. glareolus* inhabiting forests. A. *lemminus* demonstrated an increase in expression of genes responsible for muscle growth and for *Red/Ox* regulation, in particular electron transport chains. The main difference between cryptic species was associated with activation of immune processes in *L. raddei.*

Our results demonstrate how voles adapt to mountain habitat and how the expression of key genes for living in low-oxygen environments gradually increases. Also differences in expression profiles of cryptic species show various expressions of genes involved in immune processes. Our research expands the understanding of voles' adaptation and radiation and outlines future trends for exploring.

This work was supported by Russian Science Foundation grant N 19-74-20110.

УДК 576.5, 577.2

Uncovering molecular characteristics of cellular senescence process specific to human mesenchymal stem cells

P. Deryabin[1,2], L. Danilov[2,3]

[1] Mechanisms of cellular senescence group, Institute of Cytology of the Russian Academy of Sciences, Tikhoretsky ave. 4, 194064, St. Petersburg, Russia
[2] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[3] St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia

Senolytic strategies of selective targeting of senescent cells (that are known to be the major factor mediating age-related tissue malfunctioning and aging progression) represent the core trend in the anti-aging therapy area. Recently, cardiac glycosides (CGs) were suggested as broad-spectrum senolytics. The present study was triggered by the experimental evidence obtained in the Mechanisms of cellular senescence research group at the INC RAS. Their research showed the failure of the CGs to display senolytic activity towards human mesenchymal stem cells (hMSCs) of various origins. Here, we aimed to identify specific molecular characteristics of stem cells' senescence underlying their resistance to CGs-mediated senolysis. To this end, we employed comparative transcriptomic analysis of senescence development in CGs–sensitive cell types and –insensitive hMSCs. According to the results, the absence of CGs-mediated senolysis might be mediated by effective K+ cellular import supporting ionic balance and increased apoptosis resistance in senescent hMSCs. At the same time, senescence of CS-sensitive cell lines appeared to acquire pro-apoptotic gene expression profiles. This reveals that apoptosis resistance is not a general feature of senescent cells as previously postulated. Importantly, only apoptosis-prone senescent cells could be effectively cleared by CGs. Thus, we can speculate that preferential killing of senescent cells within senolytic approaches might depend on whether senescent cells are indeed apoptosis-resistant compared to their proliferating counterparts.

Raw and processed data generated in the study are deposited in the GEO database: GSE160702. Detailed description of the analysis is available at the GitHub repository: PavelDeryabin/BI_project_1st_semester. Study pre-print at bioRxiv.

This study was funded by the Russian Science Foundation (# 19-74-10038).

УДК 57.08

Simulation of DNA-probe and DNA-target hybridization in Tyramide-FISH method

A. Ermolaev[1], A. Ilin[2]

[1] Russian State Agrarian University, Moscow Timiryazev Agricultural Academy,
Timiryazevskaya st. 49, 127550, Moscow, Russia
[2] Skolkovo Institute of Science and Technology, Bolshoy Bulvar 30, 143026, Moscow, Russia

A fluorescence *in situ* hybridization (FISH) method was developed for visualizing a specific DNA sequence on a physical chromosome. However, the FISH sensitivity for detection of single copy DNA sequences was limited, especially for highly compacted plant chromosomes. An ultra-sensitive method termed Tyramide-FISH was adapted for plant cytogenetics.

Tyramide-FISH is mostly used for visualization of small targets such as genes or marker loci. In most cases genes are included in gene families and it is complicated to design a probe to visualize only one member of a multigenic family. The Tyramide-FISH method has shown that including an intron in a probe allows one to visualize a specific gene from a multigenic family. In our work we tried to figure out the influence of different introns on hybridization specificity.

Tyramide-FISH method specificity of hybridization is defined by a stringency that limits the percentage of matches and mismatches between probe and target nucleic acid that are allowed to occur without the double helix hybrid dissolution. The most commonly used stringency is 80%. At this stringency a hybrid with 80% bases or more along the probe-target hybrid being complementary and 20% or less mismatched, will remain stable. Hybrid molecules with 80% or less homology do not form or dissociate immediately.

In this project we created a simulation of the hybridization process and calculated a number of different outcomes using the Monte-Carlo approach. The simulation showed that even a single intron fragment in a probe leads to a decrease of hybridization event number with mutated target without influencing hybridization event number with real target. However, this result has not been statistically significant and the simulation method will be improved to get better results.

УДК 004.94, 575.1, 575.8

Numerical Monte-Carlo solution to the system of stochastic differential equations of population dynamics

A. Ershov, Y. Belousov

Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia

Mathematical modelling is widely used to study biological systems. One possible approach is using stochastic differential equations (SDEs) theory and Itô's formula. In this project, several modelling problems were solved with the Monte Carlo method implemented in the Python programming language. These problems include:

- modelling Poisson and Weiner stochastic process and Brownian bridge;
- modelling Itô's integral and approximate calculation of its expectations;
- approximate solutions to Itô's SDE.

The aim of this project was to construct a stochastic differential equation model for the particular problem from population biology. Two competing populations were considered and an Itô's SDE model for this process was constructed assuming known death and birth probabilities in the process over a small period of time. Then the Monte Carlo method was implemented in Python to obtain a numerical solution. This solution was compared with the solution to a corresponding deterministic differential equation model. Several results were found not exactly matching. The SDE model provided different mean population size and mean extinction time estimations and even a new possible end-state. It also allowed us to estimate the extinction possibility for each population. Thus, the stochastic model is shown to provide additional insight into this biological system.

УДК 575.8, 578.5

GWAS analysis of region of genome associated with severe Covid-19

D. Grechishkina[1], A. Evdokimova[1], V. Cheranev[1], L. Danilov[1,2], M. Raiko[1]

[1] *Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia*
[2] *St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia*

The main task in our project was to use the PLINK tool (version 1.9) to reproduce the previously reported genome-wide association study (GWAS) of disease severity in COVID-19 patients. The main purpose of GWAS is to search for relationships between a phenotype (for example, a disease) of an organism and a set of SNPs, deletions, insertions and other individual characteristics of the genome. Such analysis is usually performed on two groups (test and control group).

We took test data from the paper (https://doi.org/10.1002/mpr.1608) and used the PLINK tool for conducting GWAS analysis. We carried out several steps for quality control of the data (filtering of samples and variants based on missing genotype rates, checking for sex discrepancies, etc.) and prepared data for analysis. After the previous step we obtained the result of GWAS analysis by running association testing with the PLINK tool. First step was the chi-square test of association and logistic regression to reveal association between SNPs and binary outcome. Second step included three different types of correction for multiple comparison (Bonferroni, false discovery rate, permutation testing). At the last step of our analysis we generated Manhattan and QQ plots for the resulting summary statistics

The second step was the comparison of the results of GWAS between two recently published papers. The first one showed that severe COVID-19 is associated with a SNP in six genes (*SLC6A20*, *LZTFL1*, *CCR9*, *FYCO1*, *CXCR6* and *XCR1*) on chromosome 3 and short region of chromosome 9. The second study identified association of severe COVID-19 with specific genes (for example, *GOLGA8B* (chr15), *RIMBP3* (chr22), *HLA-A*, *HLA-B* and etc.). We suggest that the difference in results is connected with size and content of groups and methods of sequencing for data acquisition.

Project report is publicly available at https://github.com/EvdokimovaAO/GWAS_IB_fall_2020.

УДК 004.94, 577.2

Prediction protein sequences using machine learning approach

K. Faizullina[1], P. Popov[2]

[1] Higher School of Economics, Myasnitskaya st. 20, 101000, Moscow, Russia
[2] Skolkovo Institute of Science and Technology, Bolshoy Bulvar 30, 143026, Moscow, Russia

The problem of prediction protein sequences from known spatial structures has the application in study of chemical structures.

In this study, we use the structures from Protein Data Bank. As the processed and voxelized tensors have large sizes, we use only short proteins (up to 80 amino residues).

We implement 3D Convolutional Neural Network using framework PyTorch. The training of the model was done on Personal Computer. However, we could not obtain the significant decrease for loss function and have good prediction for each chain (more than 50% of correct sequence) due to lack of samples in dataset. To achieve results, we should rescale model for long sequences and port to computing on clusters.

УДК 004.94, 577.2

Orientations of contigs using Hi-C data

A. Fonin[1], N. Alexeev[2], P. Avdeyev[3]

[1] St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia
[2] ITMO University, Kronverksky av. 49, 197101, St. Petersburg, Russia
[3] GWU, Innovation Hall, 45085 University Drive, Suite 305, Ashburn, VA 20147, Washington, United States

Two decades ago, *de novo* assembly required participation of a lot of people and could be applied only for well-established model organisms. Nowadays, however, genome assembly is a commonly used method. Recent research shows that Hi-C signal may be a powerful tool to solve the remaining computational problems in the genome assembly. Existing scaffold algorithms which use Hi-C data deal with ordering contigs well (3D-DNA). However, they make errors determining orientation (when the positions of contigs are correct but their orientations are not). We build a probability model for estimating the likelihood of current contig orientation using Hi-C data and implement Monte Carlo Markov Chain algorithms to orient them according to this model. Due to that Hi-C signal may have distance at the whole chromosome and contain a lot of information about structure DNA, this approach may become a state-of-the-art solution for *de novo* assembly.

УДК 575.8, 579.25

Annotation of metagenomes of the microbial cellulolytic communities

G. Gladkov[1], M. Raiko[2], L. Danilov[3]

[1] *ARRIAM, Podbelsky chausse 3, 196608, St. Petersburg, Pushkin 8, Russia*
[2] *St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia*
[3] *Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia*

Previously, stable microbial communities were obtained that decompose cellulose on various substrates (straw, litter, sawdust). The communities were sequenced for the 16S SSU rRNA gene using the Illumina technology. In addition, for each of the communities, metagenomes were assembled based on the sequence by Oxford Nanopore technology. The aim of the work was to find the differences between microbial communities that metabolize cellulose on different substrates from ONT assemblies. The tasks of the work were to conduct *de novo* binning, compare the results with the data on 16S SSU rRNA and find community-specific glycoside hydrolases. In metagenomes 20 families of glycoside hydrolases with cellulolytic/hemicellulosic activity were found. In addition, 10 MAGs were binned. However, the genomes obtained through binning did not belong to the taxa in which the largest amount of glycoside hydrolases was found. Despite the strong differences in taxonomy between metagenome assemblies and 16S rRNA data, no strong differences in the representation of different families of hydroxide hydrolases were shown.

The work is supported by Russian Science Foundation grant named RSF 18-16-00073.

УДК 004.94, 577.2

Scaffolding based on Hi-C reads and Neural Networks

A. Ivanov[1], P. Avdeyev[2], N. Alexeev[1]

[1] ITMO University, Kronverksky av. 49, 197101, St. Petersburg, Russia
[2] The George Washington University, 1918 F St., NW Washington, DC 20052, United States

Genome assembly is a fundamental biological task on the way to understand the relationships between genome and phenotype. Despite being extremely important, it still faces a number of problems in decoding certain genomic regions and getting the full chromosome sequence. For example, a complete telomere-to-telomere human genome was assembled using numerous technologies and intensive manual correction only in 2020, despite a draft assembly being announced back in 2003.

Defining the correct orientation and order of contigs (scaffolding) is one of the challenges in genome assembly. To facilitate this process additional information such as Hi-C reads, which shows the number of interactions between genomic regions, can be utilized. In this project our goal was to develop a machine learning approach to automate the contigs reordering procedure based on Hi-C information. We implemented a deep neural network based on LeNet architecture to estimate linear distance between contigs. It takes as input 32x32 matrix of Hi-C interactions between the end of the one contig and the beginning of another contig and outputs the value in range [0; 1] corresponding to the estimated fraction from the maximum distance. We pretrained four models to predict distances up to 250 Kbp, 500 Kbp, 2.5 Mbp and 5 Mbp based on various bins resolutions (5 Kbp, 10 Kbp, 50 Kbp and 100 Kbp respectively). Training dataset was generated from the chm13.draft_v1.0 – complete telomere-to-telomere manually curated reconstruction of a human genome. Testing accuracy, estimated as a proportion of predictions with absolute error less than 0.1 from correct value, reached 84%. Those models are available at https://github.com/ivartb/HiC_scaffolding and can be used in an iterative manner to estimate the true distance between contigs. This computational tool can be adopted to combine the contigs into scaffolds.

УДК 004.94, 577.2

Analysis of batch effect

E. Khokhlova[1], A. Ivanova[1], B. Yegorov[1], L. Danilov[1,2], M. Raiko[1,3]

[1] Bioinformatics Institute, Kantemirovskaya str. 2A, 197342, St. Petersburg, Russia
[2] St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia
[3] Center for Algorithmic Biotechnology, St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia

The batch effect occurs when non-biological factors cause changes in the data produced by the experiment. For example, processing samples and controls separately or in different runs of a sequencer may lead to the batch effect. Such effects can lead to inaccurate conclusions when their causes are correlated with one or more outcomes of interest in an experiment.

The data in this study is obtained from a recent article (Gilad & Mizrahi-Man, 2015), which used RNA-Seq to study the differential expression of genes in different human and mouse tissues. Analyzing the research design, we noticed that only 1 of 5 experiments included both the human and the mouse samples, while the samples in the other experiments were processed separately, so it can be the main source of the batch effect. For our analysis, we first calculate FPKM values for each gene and each tissue across all samples. Using R we created the correlation heatmap that showed that the genes are clustered by species rather than by tissue. If this result is caused by real differences between tissues and not by batch effects, it is impossible to use and extrapolate data from model animals to humans. So we decided to try to correct the batch effect in our data.

First of all, we filtered and removed the reads which were mapped to the mitochondrial genome, but it did not affect the correlations between gene expression in different tissues. To test different approaches to batch effect correction, we applied ComBat and Harmony tools to our data and compared the results. The FPKM vectors processed with the Harmony clustered neither by species nor by tissue. On the other hand, FPKM values processed with the ComBat could be clustered by tissue rather than by species. This last result corresponds to the empirical observation that homologous gene regulatory networks establish the identities of homologous cell-types. The difference in the output of the batch effect correction methods can be explained by the data specificity and different sensitivity of these methods to the dispersion of the data. Based on our results, we recommend using the ComBat tool for batch effect correction on RNA-seq data.

All methods used are described in the repository: https://github.com/sashapff/batch-effect.

УДК 004.94, 579.2

Agent-based modeling of the spread of antibiotic resistance in a bacterial population

E. Kirillova[1], M. Serdakov[1,2], A. Ilin[1,3]

[1] Bioinformatics Institute, Kantemirovskaya str. 2A, 197342, St. Petersburg, Russia
[2] Peter the Great St. Petersburg Polytechnic University, Polytechnicheskaya st. 29, 195251, St. Petersburg, Russia
[3] Skolkovo Institute of Science and Technology, Bolshoy Bulvar 30, 143026, Moscow, Russia

Our project was aimed to model a system that allows to observe the dynamics of the spread of antibiotic resistance, which takes into account some parameters of the bacterial population. Firstly, our model considers the movement of bacteria towards the nutrient and its further absorption. We assumed that the speed of movement of resistant bacteria is less than that of sensitive bacteria. Moreover, sensitive bacteria in our model multiply with a higher frequency. Thus, we gave preference to sensitive bacteria in the absence of an antibiotic. These parameters are based on the fact that antibiotic resistance is an excess function for the cell when there is no antibiotic in the environment. Secondly, during the division of resistant cells a new cell may lose resistance due to a replication error. The opposite process is also possible, sensitive bacteria can become resistant as a result of horizontal gene transfer. Initially, all bacteria are sensitive and can become resistant only as a result of random mutations. Third, bacteria in our model die due to lack of nutrients, if they cannot be obtained; the bacteria die from the extinction of cellular functions with time; and also sensitive cells die under the antibiotics pressure.

As a result of three waves of antibiotic spread, the population survived due to the spread of resistance. When the concentration of the antibiotic decreased (between drug doses), the sensitive bacteria successfully regained their numbers. This result is a clear illustration of the fact that bacterial resistance is a huge problem, because even when a small number of resistant bacteria appear, the population never completely dies when exposed to an antibiotic.

УДК 004.94, 575.1, 575.8

Improve casual gene prioritization by patient's phenotype. Identification and parsing of relevant Mendelian disease articles

N. Konstantinovskiy[1], D. Smirnov[2], P. Holger[2]

[1] Bioinformatics Institute, Kantemirovskaya str. 2A, 197342, St. Petersburg, Russia
[2] Technical university of Munich, Arcisstraße 21, Munich, Germany

One of the main goals of a patient's deep phenotyping is to improve the analysis of genomic data for personalized medicine. In the context of rare diseases, deep phenotyping makes it possible to make a correct clinical diagnosis and based on previously published data on the association of a gene and phenotypes, to prioritize genes. The use of a uniform nomenclature for documenting patient phenotypes (HPO ontology) has allowed the development of a number of methods for gene prioritization. However, the effectiveness of these methods at the moment remains relatively low (AUC ~ 0.6). One of the approaches to increasing the efficiency of such methods is "Phenotype-driven gene prioritization for rare diseases", based on a dataset of gene-phenotype-mutation associations, which allows using machine learning methods and algorithms of graph theory to solve this problem.

One of the most important tasks in building accurate models of gene prioritization is the collection of the necessary data on the gene-phenotype-mutation association. In the field of machine learning, the amount of data plays a key role in increasing the efficiency of models, so the data collection stage is extremely important for further model creation. The existing datasets for gene-phenotype-mutation association are insufficient to create accurate machine learning models.

To solve this problem, we decided to develop a method for detecting relevant articles from the PubMed database, from which the gene-phenotype association can be extracted. This project was aimed to create a classifier of articles from the PubMed database to identify articles on Mendelian diseases. To create the classifier, we selected genes for which there is an association with a genetic disease and parsed the OMIM and PubMed databases to obtain articles on these genes. Based on this data, we trained the RoBERTa neural network. As a result, we got a classifier, which can be used to annotate the entire PubMed database and find the necessary articles for further analysis.

УДК 004.94, 577.2

Study and improvement of label propagation binning algorithm

R. Kruglikov[1], G. Ginzburg[2], A. Korobeinikov[3], I. Tolstoganov[3]

[1] Lomonosov Moscow State University, Leninskie Gori 1, 119991, Moscow, Russia
[2] St. Petersburg Academic University, Khlopina st. 8/3A, 194021, St. Petersburg, Russia
[3] St. Petersburg State University Center for Algorithmic Biotechnology, 6th V.O. line 11/21, 199004, St. Petersburg, Russia

Metagenomic binning is a process of contig clusterization which is commonly used to form metagenome assembled genomes (MAGs). There is a big variety of tools designed for that purpose. Usually these algorithms are based on contig information, such as tetranucleotide composition, length and coverage, or perform a reference-based approach. Label propagation algorithm is a novel approach that uses connectivity and coverage information from an assembly graph to improve a binning obtained from another tool. Slightly different versions of the algorithm are currently implemented as GraphBin and GraphBin2. The goal of this project is to explore the details of inner workings of GraphBin2 and attempt to improve it.

In order to evaluate GraphBin2 we gathered intermediate binning results with different parameters on 2 different synthetic datasets (MBARC26 and SYNTH64). Binning evaluation and comparison were performed with CAMI AMBER software. We show that GraphBin2 can work with very small contigs unlike other binners and bin almost all (>99%) non-isolated contigs. However, bins created by GraphBin2 have low accuracy due to usage of contigs with bad quality. High L50 of assembly also has a negative impact on binning purity. We found out that by skipping refinement stage and setting a threshold for minimum contig length and coverage allows GraphBin2 to enrich existing bins with accuracy not significantly less then that of initial binning.

УДК 577.2

Analysis of differential gene expression of non-model hydroid polyp *Dynamena pumila*

D. Kupaeva[1,2], S. Kremnyov[2,3], L. Danilov[1,4]

[1] Bioinformatics Institute, Kantemirovskaya str. 2A, 197342, St. Petersburg, Russia
[2] N.K. Koltsov Institute of Developmental Biology, Laboratory of evolution of morphogenesis, Vavilova str. 26, 119334, Moscow, Russia
[3] Moscow State University, Biological faculty, Department of embryology, Leninskie Gory 1-12, 119991, Moscow, Russia
[4] St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia

During the development of an organism, various molecular patterns are formed that determine the processes of cell differentiation and the formation of organ systems. The patterns responsible for development are quite conservative among different groups of organisms, and this is of additional interest, as the same signaling pathways can lead to the development of a complex organism structure in Hydrozoa.

Growth and morphogenesis of hydroid polyps happen with the help of a specialized organ - the growth tip. In the process of morphogenesis it lengthens and forms a shape of colony. In our study, we analyzed the differential gene expression between the shoot tip and hydrant. RNA-seq libraries of the shoot tip and hydrant were aligned on the reference transcriptome (unpublished data) and transcript expression was analysed and compared using Deseq2.

We have discovered a difference between expressions of WNT paralogs between the tip and the hydrant. Previously, we experimentally established that cWNT pathway switches the colony growth pattern. However, according to current data, the expression of *WNT3* is at a very low level in the tip. We also found that *WNT5a*, *WNT7* and others are expressed only in the hydrant, which may be important in the formation of the tip patterning.

Another interesting finding was the difference in housekeeping gene expression. For example, the expression of the *rpl18* and *btub* genes differs significantly, while *ef1a*, *gapdh*, and *syx* are expressed at the same level in both tip and hydrant. Although the cellular basis of this requires further study, we can assume that it is due to the unique cellular organization of the growth apex of hydroid polyps. It also confirms the need for more careful planning of experiments on non-model species.

All used methods and code are described and available in the following repository: https://github.com/kupaeva/Dynamena_DGE.

УДК 577.2

Study of transcriptome during intense exercises at high altitude

E. Chernyavskaya[1], D. Litvinov[2], Y. Barbitov[3], A. Maslova[4]

[1] St. Petersburg State Pediatric Medical University, Litovskaya 2, 194100, St. Petersburg, Russia
[2] Lomonosov Moscow State University, 1 Leninskiye gori, 119991, Moscow, Russia
[3] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[4] Center of medicine and genetics CERBALB, Bolshoi prospekt Vasilevskogo ostrova 90, 199106, St. Petersburg, Russia

This project aimed to study the differential gene expression in 19 sportsmen during physical and psychological stress before and after running in extreme high altitude conditions (2450-3450 m, Elbrus m.) and compare the expression profiles to the "start" point before arrival at the competition (Saint Petersburg).

We processed RNA-seq data of 19 sportsmen in 3 condition types, 57 independent RNA-Seq samples in total. Reads were mapped to the human reference genome (hg38) using STAR. The genes and isoforms expression count was conducted with RSEM. The gene differential expression analysis was performed using DESeq2. We also did two parallel analyses using gene expression data and isoforms expression data. There were fewer differential expressed isoforms compared to genes, so we concluded that using isoforms data can give better resolution.

We found that second and third conditions (before and after altitude exercise) differ less than other pairs of conditions, suggesting that altitude adaptation and not exercise shows the highest impact on the quantity of differential expressed genes.

Then we used Molecular Signatures Database (MSigDB) to determine functional groups in lists of differentially expressed genes. We found about 25 groups of genes, among which the most interesting was the group responsible for neurodegenerative diseases. It is well known that organisms can take energy from the brain during stress and high physical activity so it may be the reason for this result, but we are going to check this group in detail in the future. All our results are in the GitHub repository: https://github.com/Kate-Cher/Skyrunners.

УДК 577.2

Comprehensive analysis of shared genetic architecture between neurological and psychiatric disorders

D. Nikanorova[1], L. Protsenko[2], K. Senkevich[3,4]

[1] *St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia*
[2] *Moscow Institute of Physics and Technology (National Research University), 9 Institutskiy per., 141701,Dolgoprudny, Russia*
[3] *Montreal Neurological Institute, McGill University, QC, H3A 1A1, Montréal, Canada*
[4] *Department of Neurology and neurosurgery, McGill University, QC, H3A 0G4, Montréal, Canada*

Epidemiological studies suggested possible associations between neurological and psychiatric traits. By studying shared genetic architecture we can find common biological pathways. The aim of the current project was to find shared biological pathways by performing correlation analysis between neurological and psychiatric diseases based on the results of genome-wide association analysis (GWAS) and transcriptomic association analysis (TWAS/UTMOST).

GWAS summary statistics data were collected using the GWAS catalog and Psychiatric Genomics Consortium. Genetic correlation between each pair of traits was estimated using linkage disequilibrium score regression (LDSC). In order to identify significant associations between complex traits and gene expression we applied TWAS/FUSION (transcriptome-wide association study) software [1]. The RHOGE package with default parameters was applied to calculate correlation between each pair of disorders [1, 2]. Using UTMOST (Unified Test For Molecular Signatures), at the first step we imputed the level of gene expression according to GWAS summary statistics data and GTEx reference data. At the second step we tested the hypothesis about the association between trait of interest and the imputed expression of each gene in 44 tissues. At the last step, we performed a cross-tissue test for each gene to summarize all the statistics into a powerful metric and get the list of genes significantly associated with disease.

We have selected 21 traits (9 neurological and 12 psychiatric). High positive gene expression correlation was shown between Alzheimer’s disease (AD) and Lewi body dementia (DLB) as well as between Parkinson’s disease (PD) and REM sleep behavior disorder. High negative correlation was found between headache and obsessive-compulsive disorder, major depressive disorder and post-traumatic syndrome.

We have found a number of overlapping genes and highlighted some common pathways. *APOE* and *APOC2* genes shared between AD and DLB take part in a statin pathway, 4 genes associated with PD and AD play a role in neurotransmitter emission, a gene related to AN and PD contributes to the regeneration of skeletal muscle.

In our study, we enhance current knowledge of shared biological pathways between neurological and psychiatric traits. Identified common genes and biological pathways could be considered as novel drug targets.

References

1. Gusev, Alexander, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W. J. H. Penninx, Rick Jansen, et al. Integrative Approaches for Large-Scale Transcriptome-Wide Association Studies. Nature Genetics, 2016, 48(3), 245–252.
2. Mancuso, Nicholas, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. American Journal of Human Genetics, 2017, 100(3), 473–487.

УДК 577.2

Assembly and analysis of *Preeria caryophylla* genome

D. Panshin[1], Y. Yakovleva[2]

[1] *Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia*
[2] *St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia*

Intranuclear symbionts of ciliates were described as early as 1890, but their active study began only at the end of the last century. These bacteria belong to the Rickettsiales and Holosporales orders. Knowledge of their biodiversity is broad enough, but few genomic data of the Holosorales order are available, primarily due to the specific habitat of bacteria. In this project we assembled the genome of one of these bacteria, *Preeria caryophylla*, with reads from Oxford Nanopore (mean genome coverage ~10x) and Illumina technologies (mean genome coverage ~68x). The detailed description of software and code used in the work can be found in the following repository (https://github.com/DaniilPanshin/Preeria_caryophylla-.git).

We managed to obtain a full genome, with a total length of contigs = 831,190 nucleotides and 874 complete genes. After assembly of the genome, analysis of protein orthogroups was carried out using the Proteinortho program. 21 hypothetical proteins were found to be specific to the assembled genome of *Preeria caryophylla*. The results of analysis of orthogroups are consistent with the results obtained by Garushyants (Garushyants et al.,2018). Further pangenomic analysis of the representatives of the Holosporales order is indispensable for getting further insights into their biology.

УДК 004.94, 577.2

Improve casual gene prioritization by patient's phenotype

A. Primak[1,2], D. Smirnov[3], P. Holger[3]

[1] Lomonosov Moscow State University, Leninskie Gory 1, 119992, Moscow, Russia
[2] Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[3] Technical University of Munich, Arcisstrabe 21, Munich, Germany

It is estimated that about 6% of the human population is affected by a rare disease, of which presumably ~80% have a genetic cause. Establishing a genetic diagnosis of a rare disease, namely identification of disease causing genes and variants, can improve disease management and therapy. Phenotypic analysis along with DNA sequencing are widely established as a first tier investigation step. It is evident that systematic and automated phenotypic analysis provides significant improvement in genetic diagnostics especially for the time of the analysis. However, the power of phenotype data to predict disease-causing genes is quite low, especially in terms of specificity, due to a large number of candidate genes. Common approach to reduce the number of candidates is a restriction to the genes, harbouring qualifying rare variants.

Here we attempted to investigate the power of tissue-specific gene expression profiles to improve phenotype guided analysis. We focused on the basic dependances, i.e. dependance between involvement of a gene in any disease and its expression specificity at a tissue, the second one is between tissue specificity of gene expression and specificity of phenotype manifestation, and the same questions for dependances when using data of expression at RNA level.

To solve this problem, we wrote code in R. We studied the tissue specificity of known gene expression. We used tissue specificity scores of RNAs and proteins taken from GTeX (V8) data and the tissue specificity classes determined in (Rao, A. et al, 2018). These are tissue specific, tissue enriched, house-keeping and other genes. We used enrichment analysis of proteomics and RNA-Seq datasets and found out that genes causing diseases (disease genes) statistically significantly relate to tissue enriched class and are depleted for the "house-keeping" expression profile across tissues. Then we also studied whether the tissue specificity of expression can predict tissue specificity of phenotype manifestation. For this case we mapped HPO phenotypes from the third (related to organ systems) level to tissues. The enrichment analysis showed that if a gene expresses tissue specifically then the phenotype is also related to tissue specific class (the genes which cause tissue abnormalities statistically more frequently refer to tissue specific genes). On RNA level, affected tissues showed an enrichment not only for "tissue specific" expression pattern, but also for HK and tissue-enriched.

To sum up we found that

- disease-causing genes are depleted for the “house-keeping” expression profile across tissues. This could be potentially explained by compensatory mechanisms of other genes or high intolerance to pathogenic variation, that needs further investigation
- disease-causing genes were enriched for “tissue enriched” and “other” expression profiles
- defects in the tissue specific genes lead to the abnormalities of the corresponding tissues

Tissue specificity expression profiles show promising results to improve phenotype-based gene prioritization, however requires more investigation to be implemented in the clinical setting.

УДК 579.25

Bacterial genome assembly and decontamination

A.A. Rybina[1], L.G. Danilov[2,3], M.P. Raiko[2,4]

[1] Skolkovo Institute of Science and Technology, Bolshoy Bulvar 30, 143026, Moscow, Russia
[2] Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[3] St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia
[4] Center for Algorithmic Biotechnology, St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia

Our laboratory obtained the sample of *Escherichia coli* str. Nissle 1917 which happened to be contaminated. Two types of colonies morphologically distinct by size were produced: small and big. We assumed that the first (referred further as "Nissle Small", or simply "NS") were formed only by *E. coli* str. Nissle 1917, while the second ("Nissle Big", or "NB") consisted of *E. coli* str. Nissle 1917 and some other bacterial contaminants. Both colonies were subjected to sequencing using Illumina in paired-end mode with the read length of 75 nt, generating two sequencing samples: NB and NS.

In our work, we aimed to perform *de novo* assembly and determine taxonomy classification of contaminants for two sequencing samples obtained from the *E. coli* str. Nissle 1917 colonies. According to taxonomy classification via Kraken v1.1.1 and 16S rRNA gene homology search, sample Nissle Big was a metagenome consisting of bacteria from the species *E. coli* (including *E. coli* str. Nissle 1917) and the species *Bacillus cereus*. Using SPAdes v3.13.1, from the sample Nissle Small we obtained 4,977,707 bp draft assembly which was 4,977,707 bp long and consisted of 66 contigs with the N50 value of 205,763. At 50.53 %, the GC content is similar to that of the *E. coli* Nissle 1917 genomes available at NCBI. PGAP annotation pipeline revealed 4,773 genes, among them 4,559 are protein-coding, three complete rRNAs (2 5S and 1 23S), 4 partial rRNAs (all 16S) and 58 tRNAs. Draft assembly covered the complete genome of *E. coli* str. Nissle 1917 (GenBank assembly accession: GCA_003546975.1) by 98%. These results suggest that the studied small colonies indeed belong to *E. coli* str. Nissle 1917.

This study was funded by the Russian Science Foundation (# 18-14-00358).

УДК 577.2

Search for molecular markers - predictors of a positive response to immunotherapeutic treatment according to single-cell RNA sequencing

N. Sharaev[1], S. Tikhomirov[2], V. Zhernovkov[3]

[1] *Skolkovo Institute of Science and Technology, Bolshoy Bulvar 30, 143026, Moscow, Russia*
[2] *Moscow Institute of Physics and Technology, Institutskiy per. 9, 14170, Dolgoprudny, Moscow region, Russia*
[3] *University College Dublin, Dublin 4, Belfield, Ireland*

The aim of the study was to compare data on gene expression in people who respond to checkpoint immunotherapy and those who do not. Immunotherapy is a new approach in the treatment of malignant tumors, which consists in activating the patient's own immune response. The human immune system has self-regulation mechanisms, one of which is checkpoints, membrane molecules on the surface of immune cells that inhibit the activity of the immune system. Tumor cells use this innate self-control system against the body, activating it under inappropriate conditions. Thus, by inhibiting checkpoint activity, it is possible to allow the immune system to develop an anti-tumor response. However, for the successful use of checkpoint inhibitors, it is necessary that tumor cells use this pathway to avoid the immune response, otherwise the therapy will not give results. To decide on the need for such therapy, tests are currently being carried out for the presence of various markers in patients that positively affect the outcome of therapy. Currently, there is an active search for such markers to increase the success of immune therapies. People who respond to immunotherapy will have changes in the expression of genes involved in regulating the immune system. A change in the expression of the *HAVCR2* and *CTLA4* genes, which are genes for checkpoint proteins, was found in the respondents. Differences were also found for checkpoint ligands and tumor necrosis factors: LGALS9, CD274, CD80 / CD86, TNFSF9.

References

1. Yuanxin Wang, Ruiping Wang, Shaojun Zhang, Shumei Song, Changying Jiang, Guangchun Han, Michael Wang, Jaffer Ajani, Andy Futreal, Linghua Wang, iTALK: an R Package to Characterize and Illustrate Intercellular Communication, bioRxiv 507871.

УДК 004.94, 577.2

Detecting novel molecular events in proteomics data for genetic diagnostics

A. Sobolev[1], D. Smirnov[2], P. Holger[2]

[1] Bioinformatics Institute, Kantemirovskaya str. 2A, 197342, St. Petersburg, Russia
[2] Technical university of Munich, Arcisstraße 21, Munich, Germany

Usage of transcriptomics and proteomics in diagnostics of rare Mendelian disorders is growing rapidly and becomes a part of standard clinical practice. For prioritization of the genes, mutations in which lead to disease, three main approaches can be used in RNA-sequencing: detection of aberrant gene expression, aberrant splicing and mono-allele expression. For now, proteomics is used only as addition to the first approach. However, there is evidence that proteomics data can be used to predict expression levels of protein complexes and represent alterations in protein-protein interactions.

This project was aimed to improve diagnostics of Mendelian disorders via proteomics. To achieve that goal, we detected protein outliers in proteomics data using two different methods - LIMMA and PROTRIDER - and mapped protein complexes' database CORUM to these data. Then we used GSEA to identify enriched complexes and their function. Finally, we assessed the sensitivity and accuracy of the different aberrant expression detection methods. As a result, an optimal combination of aforementioned methods and sorting algorithms in GSEA was determined using benchmarking results and a pipeline module performing this analysis was implemented.

УДК 575.8, 579.25

Hidden biodiversity: search for uncultured protists in metagenomes

I. Sonets[1], I. Pyankov[2], Y. Yakovleva[2], M. Rayko[2]

[1] *Institute of Gene Biology, Vavilova st. 34/5, 119334, Moscow, Russia*
[2] *St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia*

Metagenomics allows to reveal the hidden diversity of uncultivated organisms, including unicellular creatures. There are groups of protists that are quite difficult to isolate from the environment and even more difficult to cultivate. In particular, the Microsporidia group is poorly investigated; hence, in our work we narrowed the focus to this group. For our study we selected 37 wastewater metagenome assemblies and 15 insect guts metagenome assemblies. To find new species, a new pipeline was developed and tested on *Enterospora canceri* to prove the concept.

Metagenomes were downloaded from NCBI using Biopython. rRNA genes were predicted using Barrnap. Results were filtered and only 18S rRNA genes were retained. Sequences were aligned and taxonomically annotated using SINA and SILVA SSU Ref NR99 database. All Dikarya and Mucoromycota-related sequences were deleted. Contigs were realigned using MAFFT. Phylogenetic analysis was performed using FastTree, and plots were made using ETE Toolkit for Python. BLAST was conducted to verify our findings.

As a result, we found 31 new sequences that belong to the Ciliophora and *Amoebozoa* supergroups of *Eukaryota*, and also sequences that are associated with *Bacteria*. At least 1 representative of Fungi, which is a member of *Opistokonta* superphylum, was found. In addition, we found some sequences belonging to potentially uncultured spcies. Unfortunately, no new *Microsporidia* genus/species were found. There are 2 possible explanations:

1. Small amount of analysed metagenomes;
2. Features of sample isolation.

To sum up, we developed a novel pipeline that allows us to find new species of protists. See GitHub repository (https://github.com/ISonets/BI_2020_metagenomics_project) for more details.

УДК 577.2

Statistical estimation of von Willebrand factor exocytosis from human endothelial cells

A. Tsitrina[1], P. Avdonin[1], L. Danilov[2]

[1] N.K. Koltsov Institute of Developmental Biology, Laboratory of evolution of morphogenesis, Vavilova str. 26, 119334, Moscow, Russia
[2] Bioinformatics Institute, Kantemirovskaya str. 2A, 197342, St. Petersburg, Russia

Endothelial cells (EC) line up the surface of blood and lymphatic vessels and regulate many aspects of human body physiology. The main functions of endothelial cells are the control of blood clotting and von Willebrand factor (vWF) is one of the major proteins in this reaction. Upon specific stimulation EC releases vWF via exocytosis from specific storage organelles – Weibell-Pallade bodies. After exocytosis, vWF can form multimeric structures called “strings” on the surface of EC. These structures activate platelet aggregation and initiate clotting formation. Recently, reactive oxygen species (ROS) were recognized as second messengers together with Ca^{2+} and cAMP. Of all ROS, H_2O_2 is the best candidate for such a role due to its molecular properties. H_2O_2 is a mild relatively stable oxidant which is highly soluble in lipids. There are two main sources of cellular H_2O_2: superoxide dismutase and NADPH-oxidase NOX4, extracellularly H_2O_2 can be generated during activation of platelets, monocytes and neutrophils. The effect of H_2O_2 on vWF release from EC is unknown.

In this work, we aimed to estimate the effect of H_2O_2 exposure on vWF release from EC in comparison with Histamine and Trombine, well known agonists of vWF secretion. All experiments were done on HUVECs – human umbilical cord endothelial cells. Briefly, cells were seeded in 48-well plate and cultured before 100% confluence and 2 days more. Then, cells were stimulated by 100 uM of H_2O_2. Histamine (100 uM) and Trombin (10 uM) were used as positive control, equal volume of phosphate-buffered saline were added as negative control. After 20 min of incubation, cells were fixed and stained by specific antibody to vWF, wheat germ agglutinin for cell borders detection and by Hoechst 33352 for nuclei detection. 25 fields of view were taken from each well, each experimental group was present as 3 independent wells. Images were segmented and fluorescence parameters were measured by CellProfiler 4.07 software. Data was exported as a CSV file and processed by Rstudio 3.6.2.

2 datasets which describe Cells and their vWF-positive structures were generated by image processing software CellProfiler with a user-defined algorithm. Final datasets contained ≈ 29000 cells and 6500 strings after outlier removing and data filtration. For estimation of statistical difference between our groups (control, H2O2, Histamine and Trombine) we choose 4 parameters in dataset Cells and 3 parameters from dataset Strings for analysis. Kruscall-Wallis

test with Dunn post hoc test were used for estimation of difference between groups due to non-normal distribution of all variables in both datasets. Epsilon-squared criterion was used for estimation of size effect. In all cases epsilon-squared were very close to zero (0.01-0.1), so even if the difference was statistically significant the difference between groups was negligible or very weak. So, manually choosing variables did not show a significant result, so that is why we make a random forest classification of our data. Before classification, we add a new factor variable with 2 levels(stimulated/unstimulated) in all our datasets and split them for 3 parts (test, train and validation sets) in ratio 0.25:0.25:0.5. Number of training variables for each iteration were determined automatically (for Strings) and manually for Cells, the number of trees were 500 for Cells and 1000 for Strings. OOB estimation for Cells random forest was 15.11%, AUC = 0.72. According to the meanDecreaseGini index, the most important variable for Cells classification was Std of vWF Intensity. OOB estimation for String random forest was 30.24%, AUC = 0.67 and the most important variable for classification was again Std of vWF intensity.

Based on random forest classification, H_2O_2 treated cells and structures were classified as stimulated, and their characteristics were more close to Histamine or Trombine treated cells.

УДК 004.94, 577.2

Lipophilicity prediction with graph convolutions and molecular substructures representation

E. Vlasova[1], A. Alenicheva[2], N. Lukashina[2]

[1] ITMO University, Kronverkskiy pr. 49A, 197101, St. Petersburg, Russia
[2] JetBrains Research, Primorsky ave. 68, 70197374, St. Petersburg, Russia

Lipophilicity is one of the factors determining the permeability of the cell membrane to a drug molecule. Hence, accurate lipophilicity prediction is an essential step in the development of drugs.

Earlier we introduced a novel approach to encoding additional graph information by extracting molecular substructures[1]. By adding a set of generalized atomic features of substructures to an established Direct Message Passing Neural Network[2] we were able to achieve a new state-of-the-art result in predicting the two main lipophilicity coefficients, namely logP and logD descriptors.

We further improved our approach, StructGNN, by adding the edges features to substructures encoder of a molecular graph and used the graph convolutional neural network (NeuralFingerprints[3] and WeaveNet[4]) approach to improve the embeddings. The WeaveStructGNN gave us a new state-of-the-art result in predicting the logP descriptor.

In addition to the previous approach, we also improved the substructures representation itself in a couple of ways. The first improvement was to add the symmetry feature based on the atom ranking, where two atoms are to be in one equivalence class if the molecule atoms can be enumerated in the same way starting with any of these atoms. Another implemented approach is adding the distance features to differentiate molecules with similar substructures representation, such as *para-xylene* and *ortho-xylene*. The last improvement is the set of features encoding the count of atoms with each hybridization type in a substructure (*s, sp, sp2, sp3, sp3d, sp3d2*). One by one these additional features did not improve the model performance, but they might be useful for the further research.

References

1. K. Yang, K. Swanson, W. Jin, et al. Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 2019, 59(8), 3370–3388.
2. N. Lukashina, A. Alenicheva, E. Vlasova, et al. Lipophilicity Prediction with Multitask Learning and Molecular Substructures Representation. arXiv:2011.12117

3. D. Duvenaud, M. Dougal, A. Jorge, et al. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In the Proceedings of Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, Canada, 2015, 7(12), 221–2223.
4. S. Kearnes, K. McCloskey, M. Berndl, et al. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des, 2016, 30, 595–608.

УДК 004.94, 578.5

Epidemic simulation of COVID-19

A. Zamalutdinov[1], A. Ilin[2]

[1] Lomonosov Moscow State University, Leninskie Gori 1, 119991, Moscow, Russia
[2] Skolkovo Institute of Science and Technology, Bolshoy Bulvar 30, 143026, Moscow, Russia

People are often confronted with different illnesses and some of them are infectious and dangerous. In order to find the balance between the health of people and economic needs and growth, modelling of such diseases is used.

Based on the SIR (susceptible-infected-removed) model, we simulated COVID-19 epidemic. To make our model more precise we added death and birth rate of the population and illness death rate as model parameters. We used COVID-19 effective reproduction number, mortality rate and illness duration obtained from several publications. However, our final model does not take into account weather changes and loss of acquired immunity (decrease in antibody levels) after some period of time.

We tested our model with different parameters of contact rates and illness duration. It demonstrates that the maximum number of infected people significantly decreases and takes place later if we reduce contact rate and illness duration. So, these factors are effective measures to reduce medical organization loading and protect economics from total quarantine and lockdown.

УДК 579.25

Analysis of metagenomic samples from black taiga and regional soil

A. Zverev[1], M. Raiko[2], L. Danilov[3]

[1] *ARRIAM, Podbelsky chausse 3, 196608, St. Petersburg, Pushkin 8, Russia*
[2] *St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia*
[3] *Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia*

In this project we analyze the metagenomes of two contrasted soils - black taiga soil (described previously as high-fertile soil) and regional regular soil. We use 2 samples of the Oxford Nanopore shotgun sequencing (N1 for black taiga soil, N2 for regional soil), and 10 16s rDNA Illumina libraries.

Complete metagenomes were assembled previously from ONT data using Flye. We use QUAST for quality control of the assemblies, MetaBAT2 and CheckM for binning, barrnap for rDNA sequence extraction, and PROKKA for annotation.

Our assemblies have the same number of contigs (about 15 000), but different lengths (61,000 b.p. for N1 and 34,000 b.p. for N2). For this reason, we were unable to directly compare the number of genes and had to normalise them first. Binning via MetaBAT2 was not successful due to high contamination level. It can be caused by the high diversity of soils and by the large number of individual genomes present in the community. At the same time, other binning approaches (for example, Maxbin) may be tried in the future. PROKKA allowed to annotate hypothetical and annotated proteins (and several hundred glycoside hydrolase genes in particular), but their numbers should be corrected.

Finally, exact 16S rRNA gene sequences, extracted from metagenome assemblies (95 for N1 and 44 for N2), all were also discovered in amplicon reads. Moreover, the taxonomic structure of the community evaluated using amplicon and metagenome data was similar on the level of Phyla.

УДК 577.2, 579.25

Search for frame shift signals in bacterial genomes

A. Petrov[1], A. Milenkin[2], I. Antonov[3]

[1] St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia
[2] Moscow Institute of Physics and Technology, Institutskiy per. 9, 141701, Dolgoprudny, Russia
[3] Institute of Bioengineering, Federal Research Centre Fundamentals of Biotechnology, Leninsky prosp. 33-2, 119071, Moscow, Russia

One of the substances necessary for the biosynthesis of chlorophyll and vitamin B12 are the Magnesium chelatase and Cobalt chelatase enzymes , consisting of three subunits and having similar functions. There are studies showing that in some organisms that synthesize vitamin B12, the cobalt chelatase contains a large cobalt chelatase subunit and two magnesium chelatase subunits. However, the genome of these organisms lacks a gene encoding the small subunits of Magnesium chelatase.

In this project, a database containing more than 1000 genomes of various bacteria containing genes of Magnesium chelatase was analyzed. Among these organisms, more than 100 signal sequences have been identified, with the help of which the reading frame is shifted and the subsequent synthesis of both medium and small subunits of Magnesium chelatase from only one gene of the medium subunit. This is possible due to the similar structure of the part of the middle subunit and the small subunit of Magnesium chelatase.

As part of the project, frame shift signals were retrieved from the database, multiple alignments were constructed for them in order to find common subsequences and specific codons responsible for the frame shift. The research revealed the most active appearance of programmed shifts in proteobacteria of the *Pseudomonas* genus.

The corresponding secondary structure was predicted for each bacterial genus separately.

Also, for the convenience of searching for genomes containing chelatase genes and analyzing the corresponding frame shift signals, attempts were made to create a convenient web interface of the existing database. The developed visualization is much more useful than just a table, because in order to make it easier to say in which biochemical pathway the key gene of the `chlD` project will take part - it is necessary to refine the entire database interface in order to have options for visualizing gene locations.

УДК 004.94, 577.2

Comparison of methods for analysis of differential gene expression

M. Komarova[1,2], K. Gainova[2], A. Kvach[2], L. Danilov[2], M. Raiko[2]

[1] Almazov National Medical Research Centre, Akkuratova st. 2, 197341, St. Petersburg, Russia
[2] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia

Differential gene expression (DGE) analysis requires that gene expression values be compared between sample groups. It means taking the normalised read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups. There are different methods for differential expression analysis such as edgeR and DESeq. The aim of this project is to compare methods for the analysis of differential gene expression.

In this project, we reanalyzed data from several articles to develop an optimal pipelines for analyzing RNA sequencing data (Arabidopsis thaliana: doi:10.1016/j.celrep.2019.11.051; doi:10.3390/genes11091057). To do this, we used several methods of aligning reads (STAR, TopHat, Salmon) and quantification methods (featureCounts, Cuffdiff, Salmon). The following raw RNA-Seq data were used:

- *Arabidopsis thaliana* - GEO datasets GSE111062; SRP133385, SRA runs SRR6767639, SRR6767640 SRR6767652, SRR6767653 (single-end, Illumina, two replicates per sample),
- *Mus musculus* - SRA dataset SRP261257 (paired-end, Illumina, three replicates per group)
- *Bugulina stolonifera*: PRJNA607082 (paired end, Illumina, three replicates per group).

First, the quality of the reads was checked using the FastQC program. The *Mus musculus* reads were of high quality and did not need trimming, others were subjected to trimming. Then RNA-seq reads for each organism were mapped to reference genome (*Mus musculus, Arabidopsis thaliana*) or de novo assembled transcriptome (*B. stolonifera*). All samples had a high percentage of aligned reads. At the next stage, counting of mapped reads was performed with different programs and data normalisation methods (featureCounts, Salmon). Received sets of DEGs were visualized in R using packages fgsea, clusterProfiler, topGo, EdgeR. As a result of the analysis, differentially expressed genes and activated signaling pathways were identified for each organism.

Comparing our results with existing data of the articles revealed significant overlapping (except bryozoan). Unfortunately during this project, we encountered some difficulties and could not achieve the initial goal. However, in the future, we plan to analyze the data by other methods and compare the results to identify the best method for RNA-sequencing analysis.

References

1. M. Carlson. org.At.tair.db: Genome wide annotation for Arabidopsis. 2019, R package version 3.8.2.
2. G. Yu, L.G. Wang, Y. Han, Q.Y. He. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 2012, 16(5), 284–287.
3. S. Ghosh, C. Chan. Analysis of RNA-Seq Data Using TopHat and Cufflinks. Methods Mol Biol, 2016, 1374, 339–61.
4. R. Patro, G. Duggal, M. Love, R. Irizarry, C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods, 2017, 14(4), 417–419.
5. M.D. Robinson, D.J. McCarthy, G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010, 26(1), 139–140.

УДК 575.8, 579.25

In search of PET-degrading enzymes: metagenomes reveal hidden homologues

D. Khaleneva[1,2], P. Dzhelali[1,2], V. Pirogov[1,3], G. Buckley[1,4], R. Shanin[1,5], L. Danilov[1,2], M. Raiko[1,6]

[1] Bioinformatics Institute, Kantemirovskaya str. 2A, 197342, St. Petersburg, Russia
[2] St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia
[3] Alferov University, Khlopina st. 8/3, 194021, St. Petersburg, Russia
[4] National Research University Higher School of Economics, Pokrovskij bul'var 11, 109028, Moscow, Russia
[5] MIREA - Russian Technological University, Vernadskogo av. 78, 119454, Moscow, Russia
[6] Center for Algorithmic Biotechnology, St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia

Polyethylene terephthalate (PET) is widely used all over the world in plastic products and its accumulation in the environment has become a global problem. Recently, various studies have been carried out to develop an efficient system for PET recycling.

Ideonella sakaiensis is a bacteria which, when grown on PET, produces two enzymes, PETase and MHETase, capable of hydrolyzing PET to produce terephthalic acid and ethylene glycol.

Our goal was to find enzymes that could potentially participate in PET processing. We reproduced the phylogenetic analysis of sequences homologous to PETase and MHETase, and then expanded it by adding information on metagenomic data.

We performed phylogenetic analysis using MEGA X software [1] and it is consistent with the earlier studies of the phylogenetic position of PETase and MHETase [2, 3]. Our studies have shown that PETase and MHETase homologs exist independently of each other in some classes of bacteria (*Actinobacteria, Cytophagales* - PETase homologues, MHETase homologues). This may indicate different evolutionary pathways for both enzymes.

Metagenomic analysis revealed a number of metagenomes with high similarity to the *Ideonella sakaiensis* enzymes, which were collected mainly as oceanic samples or compost samples. Oceanic samples of metagenomic data were collected during the expeditions of the Craig Venter Institute, and are of great interest because they contain many previously undescribed species, which may include organisms capable of processing polycarbonate chains. Similar data were independently obtained recently, and can be combined with our results to increase further known amounts of potential PET-degrading bacteria [4, 5].

References

1. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution, 2018, 35(6), 1547–1549.
2. B. C. Knott, E. Erickson, M. D. Allen, J. E. Gado, R. Graham, F. L. Kearns, et al. Characterization and engineering of a two-enzyme system for plastics depolymerization. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(41), 25476–25485.
3. S. Yoshida, K. Hiraga, T. Takehana, I. Taniguchi, H. Yamaji, Y. Maeda, et al. A bacterium that degrades and assimilates poly(ethylene terephthalate). Science, 2016, 351(6278), 1196–1199.
4. D. Danso, J. Chow, W. Streita. Plastics: Environmental and biotechnological perspectives on microbial degradation. Applied and Environmental Microbiology, 2019, 85(19), e01095–19.
5. D. Danso, C. Schmeisser, J. Chow, W. Zimmermann, R. Wei, C. Leggewie, et al. New insights into the function and global distribution of polyethylene terephthalate (PET)-degrading bacteria and enzymes in marine and terrestrial metagenomes. Applied and Environmental Microbiology, 2018, 84(8), e02773–17.

SPRING 2021

УДК 577.2

Shared biologic architecture between Parkinson's disease and inflammatory disorders

E. Chernyavskaya[1], K. Senkevich[2,3,4]

[1] St. Petersburg State Pediatric Medical University, Litovskaya st. 2, 194100, St. Petersburg, Russia
[2] Montreal Neurological Institute, McGill University, QC, H3A 1A1, Montréal, Canada
[3] Department of Neurology and neurosurgery, McGill University, QC, H3A 0G4, Montréal, Canada
[4] Pavlov First Saint Petersburg State Medical University, L`va Tolstogo st. 6-8, 197022, St. Petersburg, Russia

Parkinson's disease (PD) is the second most common neurodegenerative disease, pathologically characterized by progressive degeneration of dopaminergic neurons in the compact part of the substantia nigra. The disease etiology is associated with a complex interaction between genetic and environmental factors. There is now compelling evidence that inflammation and immunity play an important role in the pathogenesis of PD, and comorbidity between it and autoimmune diseases has been reported. But it is not known exactly what led to this association. Therefore, we want to use various methods to understand the possible cause. Thus, this project aimed to search for a common genetic architecture and specific common markers between diseases, using data from summary statistics of genome-wide association studies (GWAS).

We used GWAS summary statistics data for 6 traits (rheumatoid arthritis, multiple sclerosis, psoriasis, primary biliary cirrhosis, celiac disease, psoriasis, and Parkinson's disease) and analyzed them using two python packages: PLEIO and MTAG. We compared the results of these packages and their power and concluded that PLEIO is better for our binomial data.

We found 247 significant pleiotropic loci for our GWAS summary statistics, 7 of them were coding. Based on one of these significant variants we visualized correlations between diseases. All our results are in the GitHub repository: Parkinson-s-disease-and-inflamantion-disorders.

УДК 579.25

Identification and comparison of somatic antigen structures of symbiotic and pathogenic bacteria from *Morganellaceae* family

A. Churkina[1], A. Rybina[2], P. Kuchur[3], A. Komissarov[3]

[1] Almazov National Medical Research Centre, Akkuratova st. 2, 197341, St. Petersburg, Russia

[2] Skolkovo Institute of Science and Technology, Bolshoy Bulvar 30, 143026, Moscow, Russia

[3] Applied Genomics Laboratory, SCAMT Institute, ITMO University, Lomonosova st. 9, 191002 St. Petersburg, Russia

The study of somatic antigen structure underlies the identification and classification of Gram-negative bacteria. Somatic antigens encode surface polysaccharides and mediate interactions between the bacteria and the environment. Analysis of the functional organization of the O-antigen operons makes it possible to identify genes specific for each O-serogroup and use them both for molecular typing of strains and for identifying pathways of bacterial evolution.

The project aim is to identify and compare candidate O-antigen operons in bacteria of the Morganellaceae family with different virulence levels. It includes finding genes of O-antigen biosynthesis in the literature, annotating O-antigens genes, visualizing and comparing candidate O-antigen operons in selected bacterial species.

In this work, we created a pipeline using which, in six *Providencia* and two *Xenorhabdus* species, we detected 23 and 12 genes in variable and conserved O-antigen operons, respectively. Among them, only 7 genes (*galE*, *wxz*, *wzc*, *wza*, *ugd*, *rmlA*, *wecA*) were previously described as O-antigen ones.

Both *Providencia* and *Xenorhabdus* species have O-antigen conserved operon involved in the nucleotide (UDP- or dTDP-) sugar biosynthesis, glucosyl to lipid transfer, and O-antigen processing.

We didn't observe a correlation between lifestyle and O-antigen operon organization as initially expected. Although *Providencia* bacteria used in the study share similar virulence levels being opportunistic pathogens, their O-antigen operon significantly varies by structure. Despite different relationships towards insects (pathogenic or non-pathogenic ones), *Xenorhabdus* exhibits high similarity in O-antigen organization.

УДК 004.94, 577.2

Comparative analysis of methods for batch correction in proteomics

K.V. Danko[1,2], L.G. Danilov[1,2], A.A. Lobov[3]

[1] St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia
[2] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[3] Institute of Cytology of the Russian Academy of Sciences, Tikhoretsky ave. 4, 194064,
St. Petersburg, Russia

In this work we have analyzed proteomic data with severe technical batch effects caused by two-step biennial experimental design. Proteomic data were obtained from cells both in undifferentiated and osteogenic differentiation stages isolated from healthy donors and patients with aortic stenosis. This study aimed to define the best batch-effect correction tool in order to analyze the molecular mechanisms of osteogenic differentiation. Differential expression analysis and principal component analysis (PCA) showed an enormous batch effect driven by the year of sample collection, which was also confirmed by guided PCA (gPCA) analysis. We have tested five batch correction methods to eliminate this batch effect: ComBat, Batch-Mean Centering (BMC), Ratio A, Ratio G, and Harman. Application of all these approaches resulted in reduction of batch effect and, at the same time, the biological differences were retained. The optimal method of batch correction was chosen by comparison of gPCA and Partial Least Squares-Discriminant Analysis (PLS-DA) results. Although BMC, Ratio A and ComBat have demonstrated the best outcomes, ComBat outperformed other methods in PLS-DA analysis. Therefore, data corrected by the ComBat method was chosen for further differential expression and Gene Ontology (GO) term enrichment analysis. We have discovered that proteins involved in activation of immune response were up-regulated, whereas proteins associated with cell transporting processes were suppressed in cells in the stage of osteogenic differentiation. Here, we have eliminated the batch effect by using the ComBat method and we have also defined some molecular mechanisms of osteogenic differentiation.

УДК 577.2

Studying role of rare variants on whole genome sequencing data of Parkinson's disease (PD) patients

A. Ermolaev[1], K. Senkevich[2,3,4]

[1] Center of Molecular Biotechnology, Russian State Agrarian University, Moscow Timiryazev Agricultural Academy, Timiryazevskaya st. 49, 127550, Moscow, Russia
[2] Montreal Neurological Institute, McGill University, QC, H3A 1A1, Montréal, Canada
[3] Department of Neurology and neurosurgery, McGill University, QC, H3A 0G4, Montréal, Canada
[4] Pavlov First Saint Petersburg State Medical University, L`va Tolstogo st. 6-8, 197022, St. Petersburg, Russia

Parkinson's disease (PD) is a neurodegenerative disease characterized by the preferential degeneration of dopaminergic neurons in the substantia nigra. Genome-wide association studies have identified numerous common single nucleotide polymorphisms (SNPs) associated with PD. Altogether, these variants are estimated to explain approximately 16-36% of the heritability of PD. Lysosomal protein degradation and autophagy, including mitophagy, are strongly implicated in PD pathogenesis.

In this project we used whole-genome sequencing (WGS) data of patients from database AMP-PD (https://amp-pd.org) via Terra (https://app.terra.bio) in order to replicate data about association of rare variants of 8 lysosomal genes with Parkinson's disease. Genes *GALC, ARSA, GRN, CTSB, SCARB2, FUCA1, CTSD* and *GBA2* were selected for analysis. WGS data of patients from AMP-PD database was sample- and variant-filtered. As a result 2341 cases and 3486 controls were used for association analysis. Association analysis was performed using SKAT and SKAT-O tests for 4 groups of variants with minor allele frequency (MAF) ≤1%: all rare variants, all rare coding variants (includes non-synonymous variants), all rare functional variants (includes non-synonymous, splicing and loss-of-function variants) and all rare variants with Combined Annotation Dependent Depletion (CADD) score value >12.37. Two genes revealed a statistically significant association with PD cases. *ARSA* gene encoded arylsulfatase A showed significant association in a group of all rare variants (p-value 2.3e-04). *GALC* gene encoded galactosylceramidase showed statistically significant association in groups of all rare variants (p-value 4.3e-02), all functional variants (p-value 2.7e-02) and CADD ≥ 12.37 (p-value 2.3e-02). At the next step, we will perform genome-wide burden analysis in order to reveal other genes associated with PD.

УДК 575.8, 578.5

Identification and analysis of SARS-CoV-2-specific T-cell receptors

A. Ershov[1], A. Sobolev[1], M. Shugay[2,3]

[1] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[2]Institute of Bioorganic Chemistry RAS, Miklukho-Maklaya str. 16/10, Moscow, Russia
[3]Russian National Research Medical University, Ostrovityanova str. 1, Moscow, Russia

T-cell receptors (TCRs) are greatly variable because of VDJ-recombination, which allows distinguishing TCR clonotypes specific to certain antigen epitopes. Latest studies reveal T-cell response in COVID-19 cases. Certain TCRs that are targeting specific antigens of SARS-CoV-2 were also found. This project aims to identify a set of TCRs that can serve as COVID-19 biomarkers, develop machine learning (ML) methods that can associate individual TCR repertoires with COVID-19 status and compare existing tools for TCR clustering.

For sample classification, two TCR repertoire datasets were studied. We corrected systematic profiling method errors with V usage normalization. Frequencies of VDJ genes and CDR3 k-mers were used as features for several ML models. However, the models performed poorly when the train and test sets consisted of samples from different datasets. We further searched for most significantly different k-mers between healthy and convalescent groups. One of the datasets had no such kmers, while the top 5 significant k-mers for another came from the J segment. Such segments may distinguish convalescent samples from healthy donors.

We also compared the performance of existing TCR clustering tools (tcrdist3 and gliph2) at distinguishing SARS-CoV-2 YLQ epitopes from other virus epitopes: GLC, GIL and NLV (Influenza A, EBV and CMV, respectively) based on VDJdb data. We calculated classification metrics for multi-class and binary classification and constructed ROC and precision-recall curves. In binary classification the accuracy of both algorithms was drastically low (22% for gliph2 and 1% for tcrdist3), although areas under the curves were satisfactory in both cases.

УДК 575.8

Adding context to nonsense: analysis of sequence properties at pLoF variant sites

A. Ivanov[1], Y. Barbitoff[2]

[1] ITMO University, Kronverksky av. 49, 197101, St. Petersburg, Russia
[2] Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia

The recent progress in sequencing technologies resulted in extensive human genome sequencing and the construction of large collections of genetic variant data such as the Genome Aggregation Database (gnomAD). gnomAD contains a significant number of putative loss-of-function (pLoF) variants in disease genes, even in healthy humans. Presence of such variants may in part be explained by context-dependent nonsense suppression. To test this hypothesis, in this project we analyze the sequence context around the positions of nonsense mutations to find the possible explanation of these variants.

Firstly, we selected nonsense-mutations positions from the gnomAD v2.1.1 database lifted over to GRCh38 human reference assembly. We selected only nonsense (“stop_gained”) variants in canonical transcripts without any flags or filters provided by LOFTEE. In total, 122 101 protein-truncating variants (PTVs) were found. However, the distribution of PTVs across genes was significantly non-uniform, so only 2 variants per gene were selected for analysis (22,901 in total). We next selected highly conserved genes (LOEUF value < 0.35). We next computed the frequency of each codon in each position 90 bp up- and downstream of the variant site. As a baseline we utilized codon usage distribution based on RefSeq human CDS data from the GWU COCOPUTS (https://doi.org/10.1016/j.jmb.2019.04.021). We performed a chi-squared goodness-of-fit test to compare the observed codon usage to the COCOPUTS baseline. Position-wise p-values were adjusted for multiple testing using Benjamini–Hochberg procedure. The results show significant codon distribution differences in -1 and 1 positions relative to the mutation site. We further investigated the amino acids' distribution at these positions and found that fewer alanine-encoding codons are present before PTV sites (“-1” position). Overall, these findings suggest that nonsense suppression may play a role in shaping the PTV frequencies in highly conserved genes; at the same time, more thorough investigation of this phenomenon is needed to construct a predictive model.

УДК 004.94, 577.2

Ordering of contigs using Hi-C data

A. Ivanova[1], N. Alexeev[1], P. Avdeyev[2]

[1] ITMO University, Kronverksky av. 49, 197101, St. Petersburg, Russia
[2] GWU, Innovation Hall, 45085 University Drive, Suite 305, Ashburn, VA 20147, Washington, United States

Recent research suggests that the Hi-C signal can be a powerful tool to tackle the scaffolding problem. Scaffolding consists of three steps: orientation of contigs, ordering of contigs, and estimation of the distance between contigs. The results of the previous research projects involving contigs orientation with Hi-C data have shown promising results for using the Monte Carlo Markov chain (MCMC) for genome scaffolding. We have expanded the existing algorithm by adding the ability to reorder contigs and estimate distances between contigs using Hi-C data, which contains information about regions of the folded genome that are in physical contact with each other. Hi-C reads were used to estimate the density function of the distance between reads. This distribution was used to estimate the distances between contigs and the likelihood functions of the ordering of several contigs. Based on the obtained likelihood function, we predicted the ordering of contigs by the MCMC method. We have obtained fairly accurate results on the ordering of contigs and intend to improve the methods in the future.

УДК 575.8

Search for convergent adaptive substitutions in alpine rodents

D.A. Khaleneva[1,2], O.V. Bondareva[2,3], S.Yu. Bodrov[3], T.V. Petrova[3], N.I. Abramson[3]

[1] St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia
[2] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[3] Zoological Institute, Russian Academy of Sciences, Universitetskaya emb. 1, 199034, St. Petersburg, Russia

In new environmental conditions, organisms begin to adapt to them over time, and these adaptations can be traced at the molecular level many generations later as changes at the level of individual genes or entire systems. To determine which changes at the genome level are adaptive and which are random, it is convenient to use a model consisting of several closely related taxa that acquire adaptations to contrasting environmental conditions independently. Such a good model system is provided by rodents of the subfamily Arvicolinae, as these animals in a short time and repeatedly inhabited a variety of ecological niches from the tundra to the highlands [1,2]. Representatives of the two tribes (Clethrionomyini and Arvicolini) of the subfamily inhabit both mountainous landscapes and forest areas, adaptations occurred independently.

The aim of the study was to find nucleotide substitutions in the transcripts of mountain rodents that could potentially be associated with the process of adaptation to new environmental conditions. In the course of the work, 11 transcriptomes of mountain and forest voles (Trinity [3]) were assembled, orthological genes (Proteinortho [4]) were obtained, and the effect of selection on 251 universal single-copy orthologs was evaluated (ETE-toolkit [5]).

As a result, it was shown that 20 genes are under the influence of selection in mountain animals relative to forest animals (p-value < 0.05), which indicates the potential possibility of establishing adaptive mechanisms of rodents to live in the highlands.

The work was financially supported by RSF grant N 19-74-20110.

References

1. N.I. Abramson et al. Radiation events in the subfamily Arvicolinae (Rodentia): Evidence from nuclear genes. Dokl Biol Sci, 2009, 428(5), 713–717.
2. X. Lv et al. Climatic niche conservatism and ecological opportunity in the explosive radiation of arvicoline rodents (Arvicolinae, Cricetidae). Evolution, 2016, 70(5), 1094–1104.
3. M.G. Grabherr et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol, 2011, 29(7), 644–52.

4. M. Lechner et al. Proteinortho: detection of (co-) orthologs in large-scale analysis. BMC bioinformatics, 2011, 12(1), 124.
5. ETE 3: Reconstruction, analysis and visualization of phylogenomic data. Jaime Huerta-Cepas, Francois Serra and Peer Bork. Mol Biol Evol, 2016.

УДК 004.94, 577.2

Evaluation of DNA methylation episignatures

E. Khokhlova[1], D. Smirnov[2]

[1] Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[2] Technical University of Munich, Arcisstraße 21, 80333, München, Germany

Many genetic syndromes have unique patterns of genomic DNA methylation - EpiSignatures (disease-specific differentially methylated regions). EpiSignatures appear at the earliest stages of development and could be present in many tissues, including peripheral blood. This makes it possible to use the EpiSignatures for diagnostic testing, as well as for the interpretation of ambiguous results of genetic tests. However, to date no pipeline for discovery and validation of EpiSignatures has been published. The aim of this project was to develop a standardised workflow to generate and validate EpiSignatures. To establish the pipeline we used Illumina 450K methylation array data from 110 individuals with Mendelian disease and 125 controls available at GEO (GSE97362). The workflow includes 3 main parts: data collection and quality control, detection of EpiSignature for each disease, and training of classification algorithms. To detect EpiSignatures, 1) differentially methylated regions were identified with the limma R package, and 2) significant probes were filtered based on their pairwise correlation and predictive power. For interpretation, a Random Forest classification algorithm was implemented. Developed pipeline can be supplemented with data from patients with other diseases to expand its predictive capabilities.

УДК 004.94, 577.2

Age-related changes in methylation patterns at sites of active human DNA replication

E. Kirillova[1], N. Konstantinovskiy[1], R. Cherniatchik[2]

[1] Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[2] JetBrains Research, Primorsky ave. 68-70, 197374, St. Petersburg, Russia

Replication of eukaryotic DNA begins simultaneously at different sites on the DNA chain, such sites are called “replication origins”. According to published data, sites of replication initiation are associated with the state of chromatin and may be of interest from an epigenetic point of view. The aim of the project is to investigate methylation patterns changes in replication origins regions in people of different age groups.

The main objectives of the project are: create a Snakemake pipeline to get consistent data, develop a method for determining conservative peaks, select methylation datasets for different age groups, select methylation data for replication origin regions and assess methylation signal changes and methylation variability across studied groups. The constructed Snakemake pipeline includes several steps, the key ones of which are: download fastq files, bowtie2 alignment, bam coverage, MACS2 peak-calling. Following the implementation of a pipeline, we then developed a technique for detecting conservative peaks – we consider a peak to be conservative if it is present in at least half of the input data files. Then we carried out the procedure for expanding the peaks to include neighboring genomic regions and increase the chance of overlap between the peaks.

After that, we used the obtained conservative peaks to check the differential methylation between young and old groups in the obtained regions. As a source of methylation data, we used the RRBS data, where methylation level was measured in CpG-islands for CD14+ CD16- monocytes of two cohorts of healthy young and older donors. Using the data obtained, we carried out an analysis of differential methylation using the Mann-Whitney test with FDR level = 0.05. The gene sets linked to significantly differentially methylated regions (we link targeted replication origins of the genome with neighboring genes using GREAT) were then matched against biological pathways and Gene Ontology databases.

As a result, using the developed pipeline, we got 56 replication origins narrowPeak files. For the peak expanding procedure, we used a 3000 bp window. Then we got the following Mann-Whitney test results: 262 regions were identified, where the methylation level statistically significantly changed with age, among which in 187 regions, methylation increases with age and in 74 regions the methylation level decreases. As a result of using GREAT, we obtained 312 up-methylated genes and 140 down-methylated genes located in close proximity to differentially methylated replication origins regions. We were unable to identify any enriched signaling pathways, but we did find interesting associations with biological processes such as neurogenesis, regulation of the cell cycle, sensory perception, and regulation of blood pressure which were nominally significant but did not pass FDR threshold. In general, the results obtained demonstrate that further study of the issue is of interest, and the written pipeline for the study of regions can be used for further testing on other methylation datasets.

УДК 577.2

Analysis of transcriptome signatures of patients with heart failure and with myopathy

M. Komarova[1,2], O. Ivanova[1]

[1] Almazov National Medical Research Centre, Akkuratova st. 2, 197341, St. Petersburg, Russia
[2] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia

In heart failure (HF), functional and metabolic alterations are detected not only in cardiac muscle but also in skeletal muscle tissue. In our laboratory based on Almazov National Medical Research Centre we have conducted transcriptomic analysis of calf muscle of healthy donors and patients with heart failure. We have also found another study (DOI: 10.1161/JAHA.120.017091) where transcriptome analysis of the pectoral muscle was performed. The aim of this project was to investigate the muscle transcriptome of patients with heart failure in a published study and compare these results with the same experiment performed in our laboratory.

Firstly, we assessed the quality of the initial reads using FastQC v0.11.9, aligned them to the genome (GRCh38 human reference assembly) using STAR v2.7.3a, and calculated per-gene read counts using featureCounts v1.12.0. Differentially expressed genes were determined using a DESeq2 package in R with a pairwise comparison of conditions. Only genes with Log2FoldChange>1 and padj<0.05 were considered as differentially expressed. To find significantly enriched pathways we performed Gene Ontology (GO) term enrichment analysis (gseaGO function), fast Gene Set Enrichment Analysis (fgsea function) and enrichment analysis (enrichGO function) using R.

We reprocessed RNA-seq data from a previously published study and identified just two differentially expressed genes (DEG). Due to the small number of DEG, only gene set enrichment analysis on the whole gene list was performed. We used the gseGO and fgsea functions from clusterProfiler library in R. In both cases, we got the same up and down regulated signal pathways in patients with heart failure.The pathways that control cellular respiration and protein localisation were downregulated in patients. Comparison of healthy donors and patients from our laboratory with fgsea gave many DEGs and slightly different corresponding signaling pathways. This can be explained by the not very good choice of the control group in the article and by another muscle.

The PCA plot showed that the patients in the article are divided into two groups according to the first component. We have found that this is influenced by the left ventricular ejection fraction. The second group (HF_2) is characterized by a lower (almost critical) fraction compared to the first group (HF_1). We compared the two groups of patients and found 1062 DEGs. In this case we also did GO analysis, FGSEA analysis and enrichment analysis on DEG. We identified more differentially expressed genes when comparing healthy donors and patients. Using GO analysis and FGSEA we found that the left ventricular ejection fraction has an effect on RNA splicing, various metabolic pathways, immune response, cell adhesion and muscle contraction. Most of the signal pathways in the patient's group with very low left ventricular ejection fraction (HF_2 group) were downregulated. They included RNA splicing and cilium organization.

Thus, in the present work we have shown that the left ventricular ejection fraction has a key influence on the development of skeletal muscle atrophy in heart failure.

УДК 575.8

Analysis of evolutionary conservation of polyproline motifs

R. Kruglikov[1], M. Parr[2]

[1] *Lomonosov Moscow State University, Leninskie Gori 1, 119991, Moscow, Russia*
[2] *Technische Universität München, Wissenschaftszentrum Weihenstephan, Department of Bioinformatics, Am Staudengarten 2, 85354, Freising, Germany*

Polyproline motifs are consecutive stretches of prolines with flanking residues. They occur in proteins of prokaryotic and eukaryotic organisms. Polyproline motifs are of great interest because of their ability to stall ribosomes during translation. A consecutive stretch of prolines coded in mRNA blocks peptidyl-transferase reaction due to the high rigidity of the pyrrolidine ring. Polyproline motifs are hypothesized to play a regulatory role in co-translational folding providing structural motifs with a sufficient amount of time to fold. However, this field as well as the role of polyproline motifs remain poorly studied . The goal of this work was to explore the evolution of polyproline motifs and their possible role in the formation of secondary and tertiary protein structures.

For evaluation of polyproline motif conservation, we downloaded proteomes of 43 different *Escherichia coli* strains. All proteins were divided into non-hierarchical orthologous groups. Further analysis included pairwise alignments of proteins within groups and analysis of substitutions that affect prolines. We showed that prolines within polyproline motifs are less conserved than single prolines. Analysis of the conservation of polyproline motifs in core and accessory proteomes has demonstrated that polyproline motifs in core proteomes are more conserved than in accessory proteomes. These results suggest that there is no special evolutionary pressure preserving polyproline motifs in protein sequences, however, the difference in their conservation in core and accessory proteomes and the intriguing regularity of their distribution in protein sequences provide the space for further investigations.

УДК 579.2

Influence of microbiota on the effectiveness of immunotherapeutic drugs in the treatment of metastatic solid tumors

D. Kupaeva[1,2], K. Sogomonyan[1,3], S. Sidorenko[4]

[1] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[2] N.K. Koltsov Institute of Developmental Biology, laboratory of evolution of morphogenesis, Vavilova st. 26, 119334, Moscow, Russia
[3] St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia
[4] Children's Research and Clinical Center for Infectious Diseases of the Federal Medical and Biological Agency, Professora Popova st. 9, 197022, St. Petersburg, Russia

Tumor cells use immune checkpoint pathways to evade the host's immune system. The use of immune checkpoint inhibitors can suppress this signal. This type of therapy can be an effective strategy for the treatment of patients with solid metastatic tumors. However, the outcome of immunotherapy is difficult to predict and it is ineffective for many patients. Gut microbiota is shown to be one of factors leading to the success of therapy.

In this study we used reads of 16S rDNA obtained from 71 cancer patients. To assess the taxonomic diversity we used two different pipelines: dada2 and Qiime2. As a result of dada2 workflow, ASV were obtained, analyzed using the phyloseq package and normalized with the Deseq2. Qiime2 output is an OTU table, the data of which is already normalized. We performed alpha-diversity analysis based on the Shannon diversity index and beta-diversity analysis using Bray-Curtis distance, Jaccard index, UniFrac and rarecurve. Profiling of predictive gut microbiota was analyzed by using three different types of linear regression, and random forest methods. The dependent variables in these models were the presence or absence of progress and type of objective response.

As a result we found that the microbiota of patients with different types of response to therapy does not differ in either the alpha or beta diversity of the community. The use of machine learning methods allowed us to build a model that can predict the treatment outcome. Analysis of the most influential predictors of the model reveals about 5 taxa for which correlations with the success of therapy are already known. However, the list of predictors is not reproducible when using a different set of methods.

УДК 004.94, 577.2

Developing best practices for single-cell analysis: data integration

D. Litvinov[1], M. Serdakov[1], V. Tsvetkov[2]

[1] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[2] ImmunoMind Inc., Berkeley, United States

The aim of our project was to develop an algorithm for comparing various libraries for integrating scRNA-seq data. Sequencing of individual cells makes it possible to analyze information about individual cells, which in turn makes it possible to distinguish cells of a subpopulation with high resolution. This turns out to be necessary primarily for pathological cell lines since such cells are highly heterogeneous. Analysis of differences in the level of gene expression in individual cells makes it possible, for example, to use genes characteristic of certain diseases.

Nevertheless, the analysis of the data of such experiments is rather difficult due to the large amounts of data, as well as biological and technical batch effects, which make it difficult to determine the key factors for the analysis. Currently, there are about 500 different Python and R packages for scRNA-seq data analysis. Accordingly, there is a need to develop some "gold standard" for working with this data.

To find the optimal algorithm in this project, we have developed a tool that allows us to compare 5 different Python libraries for working with scRNA-seq data. These libraries use different algorithms to correct the batch effect: Combat and “Regress out” use linear models for batch correction, while MNN, BBKNN, and Scanorama are looking for mutual nearest neighbors in other batches for every cell. Silhouette-score with cosine distance was chosen as a metric to assess the quality of the correction.

To compare the algorithms, we used three public datasets of different biological origins (Human dendritic cells, Human pancreas and Mouse retina) with different numbers of cells (500, 15000 and 80000, respectively) and batches (2, 4 and 2 respectively). Also, in addition, we generated 3 artificial datasets with similar parameters using the SymSim R library. The creation of artificial sets with a predetermined set of data parameters is an important task for testing the algorithms, since the way the algorithm analyzes such artificial data allows one to determine the accuracy of the clustering and correction of the batch effect carried out by the algorithm. As a result, we have created a program that can apply any of the 5 tested algorithms for batch effect correction chosen by the user or run all algorithms and select a list of the most efficient ones. The tool we have developed is in the public domain and, we hope, will allow researchers in this area to facilitate the choice of a library that allows them to process a particular dataset with the highest level of accuracy.

УДК 577.2

Fine-mapping of causal variants in the *SCARB2* and *CTSB* genes in Parkinson's disease

D. Nikanorova[1], Z. Gan-Or[2]

[1] St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia
[2] McGill University, QC H3A 0G4, 845 Rue Sherbrooke Ouest, Montréal, Canada

Several important risk loci for Parkinson's disease (PD) are associated with lysosomal proteins. For instance, *GBA* encodes glucocerebrosidase, a lysosomal hydrolase responsible for degrading glycolipids. Single nucleotide variants in *GBA* are found in up to 20% of PD patients from different populations, thus being crucial risk factors for this disease. According to data from blood and brain tissue of PD patients, the enzymatic activity of glucocerebrosidase can be altered even among PD patients without variants in *GBA* itself, which implies that other factors can influence glucocerebrosidase activity. We assume that such factors may be genetic variants in related proteins: the transporter of GBA Scavenger Receptor Class B Member 2 (encoded by the *SCARB2* gene) and cathepsin B, a lysosomal protease, previously shown to be involved in α-synuclein degradation (encoded by the *CTSB* gene). (McGlinchey and Lee 2015).

This study is aimed at pinpointing most likely variants that drive the associations of *SCARB2* and *CTSB* genes with PD risk. Starting from conditional analysis via GCTA-COJO we identified 1 independent variant in each locus, assuming that a single LD-block is present in each of the loci. Then we applied two statistical fine-mapping approaches: FINEMAP and SuSiE in order to define credible sets of probable causal variants. These methods are used to perform variable selection in multiple regression and are particularly suitable for settings where some of the X variables are highly correlated, which is a key feature of GWAS, where variants are highly correlated in LD-blocks. We used publicly available data from 1000 Genomes Project as well as IPDGC-NeuroX data to calculate the LD reference panel with PLINK 1.9. Finally, we carried out colocalization analysis via COLOC and SMR tools that aim at finding the association between mutations and gene expression through detecting shared variants between GWAS and eQTL data.

Through this workflow we narrowed down a number of most associated SNPs to 14 in the *CTSB* locus and 36 in the *SCARB2* locus. Among them 5 variants are located directly in the *SCARB2* gene, and 3 - in *CTSB.* All the variants belong to introns or UTR-regions. Significant colocalization was found between SNPs in the *SCARB2* region and variants affecting expression of the genes from this locus (eQTLs) in brain and blood. For 3 genes in *SCARB2* locus (*SCARB2, CCDC158* and *FAM47E*) and for *FDFT1* gene in *CTSB* locus the colocalization was predicted to be driven by a single shared variant. Interestingly, some of the variants in *SCARB2* gene are responsible for colocalization with different eQTL studies and tissues. We can assume that they have possible pleiotropic effects in different tissues.

References

1. McGlinchey Ryan P., Jennifer C. Lee. Cysteine Cathepsins Are Essential in Lysosomal Degradation of α-Synuclein. Proceedings of the National Academy of Sciences, 2015, 112(30), 9322–9327.

УДК 577.2

Preoperative assessment of glioblastoma resectability and development of a phenotypic rating scale

D. Panshin[1], I. Babkina[2]

[1] *Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia*
[2] *St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia*

Glioblastoma is the most common and most aggressive form of brain tumor. Surgery remains the standard treatment for glioblastoma. Several rating scales have been developed to determine the degree of tumor resectability. They are based on the patient's history. However, the effectiveness of these scales has not been evaluated in real practice. We obtained a dataset containing 60 phenotypic features for 114 operated patients with glioblastoma. The aim of our work is to compare the effectiveness of the 3 proposed scales by assessing the predicted and real outcome of the operation. In case of ineffectiveness of any of the scales, we had a task to try to develop our own approach to risk assessment based on the available data. The predictive abilities of the existing scales were assessed and compared with each other using ROC-curves, as well as using the F-score. None of the scales showed an exceptional result, the best of the proposed was the scale proposed by Capellades (AUC = 0.5952, F1-score = 0.36). We next decided to apply machine learning methods in our work to construct a custom predictive scale. We trained the random forest model which showed the best results across all tested approaches (AUC = 0.6352, F1-score = 0.4). The development of an effective prognostic scale is an important task that can greatly alleviate the problems of modern medicine.

УДК 577.2

Comparative analysis of genomic assemblies, repertoire of repeats, and protein-coding genes of palm trees *Elaeis guineensis* (*pisifera*) and *Elaeis oleifera*

L. Protsenko[1], A. Andreev[1], T. Tatarinova[2]

[1] *Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia*
[2] *La Verne University, 1950 3rd St, La Verne, CA 91750, California, United States*

The aim of the project was to monitor the quality of two palms (*Elaeis pisifera* and *oleifera*) genomic assemblies.

According to BUSCO statistics, *Elaeis guineensis* (*pisifera*) genome comprises 95% busco orthologous genes meanwhile *Elaeis guineensis* reference genome has only 91%. *Elaeis oleifera* genome comprises 85% BUSCO orthologous genes hence its assembly is not absolutely completed. If aligned *pisifera* genome to a *pisifera* reference one, it gaved 93% identity. Due to the absence of any reference data for *Elaeis oleifera*, it was aligned to *Elaeis guineensis* with 22% identity. GDSL esterase/lipase 1 was a single annotated gene with the best alignements between palms. When conducting the blast tool using 89339 potentially *Elaeis guineensi*s genes as database and *Elaeis oleifera* genome as query, we received 4011 genes with alignment length of at least 1Kb and 99% identity. These results highlight the uncompleted annotation of *Elaeis guineensis* genome with mediocre quality of oleifera assembly.

According to the information that we received from our supervisor, it was potentially evidence of a PiggyBac transposable element in *Elaeis oleifera* genome. To check this hypothesis, we de-novo created a database of repetitive sequences for both palms via RepeatModeler2. Nothing like PiggyBac was found in both palms. In general, 18 classes of repetitive elements were common. To check if PiggyBac element exists but is not determined by RepeatModeler, we firstly downloaded RepeatMasker results that were provided by the supervisor (with PiggyBac inside) and extracted genomic locations with PiggyBac. Then we compared new RepeatMasker results (using RepeatModeler library) with the provided one and received Unknown elements with the same locations with PiggyBac before. To check the plausibility of both RepeatMasker results, we compare the names of the classes of transposable elements at the same positions and it corresponded with 99% accuracy. To recheck it, we blasted potentially PiggyBac sequences to NCBI and Dfam databases, but received nothing. These findings confirmed the wrongness of the hypothesis of PiggyBac potential presence in *Elaeis oleifera* genome.

УДК 577.2

Bioinformatics analysis of E3 ubiquitin ligase family

A. Shemyakina[1], I. Pyankov[2], D. Andreeva[2] , A. Petrov[2], P. Popov[3]

[1] *Scientific Center "Kurchatov Institute" Research Institute for Genetics and Selection of Industrial Microorganisms, 1-st Dorozhniy pr. 1, 117545, Moscow, Russia*
[2] *St. Petersburg State University, Universitetskaya emb. 7-9, 199034, St. Petersburg, Russia*
[3] *Skolkovo Institute of Science and Technology, Bolshoy Bulvar 30, 143026, Moscow, Russia*

Proteolysis-targeting chimeras (PROTACs) induce targeted protein degradation by the ubiquitin-proteasome system. They represent a new therapeutic modality and are the focus of great interest, however the progress is hindered by the low efficiency of protein crystallography that provides E3 ubiquitin ligases' 3D structures required in the initial steps of PROTAC development. The automated *in silico* modeling tool could assist in expanding the number of enzymes available for development of targeted protein degradation systems. Binding site prediction on E3 ligase is another challenge in PROTAC designing. In this work we built a pipeline that executes molecular modeling of target protein and detected binding site predictions for three existing E3 ubiquitin ligase structures.

The pipeline was tested on E3 ubiquitin-protein ligase TRIM69. First the search with mafft L-INS-i selected 14 homologues in SWISS-MODEL Repository and World Wide Protein Data Bank. After homologues' processing the script selected 4 templates that provide 89% coverage of target protein and have homology percent over 15. The script generated a modeled by Rosetta 3D structure and analyzed its quality across TRIM69 length with Ornate 3D CNN. The resulted mean score was less than 0.1 lower that the score for TRIM69 model generated by Robetta server.

In order to predict binding sites we ran a full-atom molecular dynamics simulation in water for three ligases (TRIM25, HUWE1, FBXW7) via GROMACS tool, predicted binding sites across the obtained trajectory using BiteNet tool, clustered the obtained predictions and finally selected the most promising ones.

УДК 575.1, 575.8

Inference of inbreeding coefficients in GADMA

S. Iliutkin[1,2], A. Sidorin[1,2], E. Noskova[3]

[1] Bioinformatics institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[2] St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia
[3] ITMO University, Kronverksky pr. 49A, 197101, St. Petersburg, Russia

GADMA is a software that implements methods for automatic inference of the joint demographic history of multiple populations from the genetic data. Demographic history is the record of the population's development in the past. Models of demographic history inferred from genetic data complement archeology and serve as null models in genome scans for selection. Inbreeding is the production of the offspring by mating of organisms that are closely related. Unaccounted inbreeding coefficients in models of demographic history may lead to the wrong inference in case of inbred populations.

GADMA is based on the two open-source packages for inferring demographic history: dadi and moments. GADMA is a command-line tool, it can take a simple file of run settings as input and generate input and code for dadi or moments. The latest version of the dadi allows to account for the inbreeding coefficients in demographic models; however, in the current version of GADMA the estimation of the inbreeding coefficient was not implemented.

We have updated GADMA and incorporated inference of inbreeding coefficients. Now users can add inbreeding as a parameter of the demographic model. We have supplemented the generator of the dadi code to include parameters of the inbreeding. We also have covered the new code with tests. Now GADMA may include inbreeding in the automatically built model and work with it. To evaluate the correct implementation of inbreeding in GADMA we used data from the study about the demographic history of American puma. GADMA was successfully launched and estimated inbreeding coefficients are close to those from the original paper.

УДК 577.2, 578.5

Search for the insertion site of human betaherpesvirus 6A into the human chromosome

I. Sonets[1], A. Popova[2], O. Goleva[3], Yu. Eismont[4], O. Glotov[4]

[1] Institute of gene Biology, Vavilova st. 34/5, 119334, Moscow, Russia
[2] Moscow State University, GSP-1, Leninskiye Gory 1, 119991, Moscow, Russia
[3] Pediatric Research and Clinical Center for Infectious Diseases, Professora Popova st. 9, 197022, St. Petersburg, Russia
[4] City Hospital № 40, Borisova st. 9, 197706, Sestroretsk, St. Petersburg, Russia

It has been shown that human betaherpesvirus 6 (HHV-6A) can integrate into the telomeres of the host cell chromosomes. The chromosomal integrated form of the virus (chiHVH-6A) occurs in the human population in up to 1.5% of cases and can be inherited from one or both parents to children. The sites of integration are not clearly defined.

The main goals of this project were to make a hybrid assembly of the human genome, presumably containing integrated HHV-6A and to search for the insertion site of HHV-6A in the assembled genome. DNA was isolated from sperm cells of an infected male subject, whose son was hospitalized with infection associated with HHV-6A and also infected with HHV-6A. Whole genome DNA sequencing was performed with MGI & Oxford Nanopore. Two *de novo* hybrid assemblies using HASLR and WENGAN algorithms were produced, but showed inadequate results. To compensate for this, Nanopore-only and MGI-only assemblies were made. MGI-based assembly made with Minia showed potential good results, but no integration event was found. QUAST results for this assembly are: GC% = 38.45, N50 = 861, L50 = 397067, largest contig = 9063, total size ~ 3.12 Gb. QUAST results for Nanopore-only assembly made with Flye are: GC% = 40.5, N50 = 144131, L50 = 443, largest contig = 542677, total size ~ 768 Mb. Using Minimap2 to align raw Nanopore reads onto the reference HHV-6A genome, we showed presence of an integration event, but exact location is still unknown. Custom pipeline was proposed for more efficient assembly and currently in WIP state. Future plans include Nanopore re-sequencing to improve coverage, as also filter out bacterial contamination, and usage of more sophisticated algorithms to improve assembly quality.

УДК 004.94, 577.2

Building gene regulatory networks with structural properties

O. Vavulov[1,2], E. Zhivkoplias[2,3]

[1] Sberbank PJSC, Vavilova st.19, 117997, Moscow, Russia
[2] Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia
[3] Science for Life Laboratory, Box 1031, 17121, Solna, Sweden

Gene regulatory network (GRN) is the model used to describe gene expression levels and their complex interconnection on the whole-organism scale. It is represented as a graph that can be inferred from expression data, but the inference quality can not be properly assessed without known true structure which is the case in most real-world tasks. The answer is to challenge inference algorithms by artificial expression data generated from the random but still plausible simulated network.

In this work, the original network generation algorithm was implemented using Python. The algorithm was designed to capture the main local structural features of the real GRNs: 1) feed-forward loops (FFL) abundance compared with randomized networks of the same structure; 2) vertex-based FFL network is the connected graph. Addressing these properties, at the first stage of generation, the whole FFL is attached until a predefined network size is reached. The classic preferential attachment algorithm is employed for network finishing afterwards. The resulting artificial GRN has a motif number distribution similar to the real GRNs' ones and captures the specific high FFL abundance. One can tune the size and sparsity of the generated network using parameters.

УДК 004.94, 577.2

Haplotype variant calling using hypergraphs approach

E. Vlasova, N. Alexeev

ITMO University, Kronverkskiy pr. 49A, 197101, St. Petersburg, Russia

Many viruses such as human immunodeficiency virus and hepatitis C virus have a very high mutation rate, which results in several haplotypes, closely related genomic variants, coexisting in the carrier's organism. Thus, identifying the haplotype sequences and their percentage in the mix is a very important task for biologists and doctors in order to create a more efficient treatment.

We present HyperHaplo, a novel method for haplotype variant calling from the NGS data which is based on the hypergraph approach. We build an allele graph, where a vertex is a single nucleotide polymorphism (SNP) and a hyperedge is a set of vertices. A hyperedge connects vertices if they appear in one read. After creating starting hyperedges, we run the merging algorithm. According to the algorithm, we merge two hyperedges if they intersect and do not conflict. The algorithm always tends to assemble the most frequent haplotype in a mix so the most frequent haplotypes are assembled mostly correct. We consider that a haplotype is assembled as soon as a hyperedge contains all the SNPs.

We used the HIV lab mix[1] dataset to test the algorithm’s behavior via the UniFrac and mean minimum distance metrics. The dataset is an Illumina sequencing of a mix of five HIV strains. The UniFrac metric is the Wasserstein distance and the mean minimum distance is a metric which looks for each predicted haplotype the closest real strain and calculates the average value over all the predicted haplotypes. Our method outperforms the recent ITMO hypergraph approach, but it is unfortunately worse than the CliqueSNV[2] approach.

References

1. Di Giallonardo, Töpfer et al. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations, NAR, 2014, 42(14), e115.
2. Sergey Knyazev et al. CliqueSNV: Scalable Reconstruction of Intra-Host Viral Populations from NGS Reads, bioRxiv 264242.

УДК 577.2

Transcriptomics and small RNAome responses to the infection by phytopathogenic fungi: an analysis in the wild model legume *Medicago truncatula* infected by *Verticillium alfalfae*

A. Zamalutdinov[1], L. Gentzbittel[2]

[1] *Lomonosov Moscow State University, Leninskie Gori 1, 119991, Moscow, Russia*
[2] *Skolkovo Institute of Science and Technology, Bolshoy Bulvar 30, 143026, Moscow, Russia*

Resistance mechanisms to *Verticillium* wilt are well-studied in tomato, cotton, and *Arabidopsis*, but much less in legume plants. However, legumes are a significant source of protein, dietary fiber, carbohydrates and dietary minerals and also used as livestock forage. Moreover, most of them have symbiotic nitrogen-fixing bacteria.

Medicago truncatula is a model for legume plants and particularly attractive for the study of plant-microbe interactions. It has a small diploid genome, is self-fertile, has a rapid generation time and its genome has been sequenced.

In our research two lines of *M. truncatula* were used: resistant and susceptible to infection with the *Verticillium* wilt. Plants of each line were cultivated in two conditions: non-inoculated and inoculated. For sequencing purposes roots of 6 plants were collected in each condition and line combination. Roots were collected before inoculation and then during the early, intermediate and late phase of infection.

In the first stage of our work we identified differentially expressed genes (DEGs) during the late phase of infection. We identified 928 DEGs between the strains (resistant or not), as well as 563 genes regulated by pathogen inoculation and 67 regulated by resistance:inoculation (genes that are regulated by inoculation in resistant line only). We conducted gene set enrichment analysis (using AgriGO) for all DEG groups with the exception of genes downregulated by pathogen inoculation (only 3 genes were identified in that group). For a group of genes downregulated upon inoculation in resistant lines, no enriched biological process terms were found. The DEGs in general are involved in transcription, biosynthesis regulation, ncRNA processing, DNA repair and many other processes. Most of the cellular components are involved in these processes.

Our results are the initial step towards understanding mechanisms of legume resistance to fungi pathogens.

BIOINFORMATICS SUMMER SCHOOL 2021

УДК 575.8

Филогения *Lumbrineridae* (*Annelida*) на основе молекулярных данных

П. Борисова, Н. Будаева

Институт океанологии им. П.П. Ширшова РАН, Москва, Россия

salixhastata@ya.ru

Lumbrineridae — семейство морских кольчатых червей с бедной внешней морфологией, но со сложно устроенным челюстным аппаратом. Детали строения челюстного аппарата имеют высокий таксономический вес, в первую очередь, на родовом уровне. *Lumbrineridae* включают 19 родов, предложенных в результате единственного филогенетического анализа на основе морфологических данных. Ввиду бедности внешней морфологии, система люмбринерид с высокой вероятностью основана на использовании гомопластических признаков и нуждается в дополнительном тестировании с использованием генетических данных. В данной работе мы реконструируем филогению семейства методом баесовского анализа, а также проверяем монофилетичность принятых в настоящее время родов на основе последовательностей участков трех генов: 16S рДНК, 18S рДНК и COI. Молекулярные данные были получены для 10 родов, для некоторых родов впервые. Нами были исследованы детали строения челюстных аппаратов, помимо световой микроскопии мы использовали компьютерную микротомографию.

Филогенетический анализ показал, что семейство *Lumbrineridae* является монофилетическим таксоном с высокой поддержкой (PP=1). Полученная реконструкция во многом противоречит общепризнанной системе семейства. Так, самый крупный и типовой род семейства *Lumbrineris* согласно нашим данным является полифелитичным. Род *Abyssoninoe* также вероятно является полифелитичным. Рода *Augeneria, Ninoe* являются монофилетичным. Было показано, что некоторые признаки морфологии челюстей, ранее рассматривавшиеся как синапоморфии крупных клад, являются гомопластическими.

УДК 004.94, 577.2

Using protein-protein interaction and gene networks to improve Connectivity Map

M. Minaeva, K. Murtazalieva, Y. Medvedeva

Moscow Institute of Physics and Technology, Institutskiy pk. 9, 141701, Dolgoprudniy, Russia

minaeva.mav@gmail.com

Ample data on drug-induced changes in transcriptomic profile could be used to identify new applications of existing drugs. One of the first systematic approaches for this purpose was Connectivity Map. It compares a request signature that represents the difference between two states of interest with the signatures induced by various interventions. However, this approach does not consider the biological importance of differentially expressed (DE) genes in the signature. We have developed an efficient approach of Connectivity Map based on topological metrics of protein-protein interaction and regulatory networks data. It deals with the biological importance of genes and can be used to predict chemicals that might induce desirable changes in transcription profile. Our tool reveals small molecules that could provoke cellular conversions. As an input we use signatures created from DE genes between two cellular states that we want to reverse or mimic. Our main hypothesis: if the signature in response to a perturbagen from the database shows a high overlap with the query signature then this substance has a high probability to cause a similar cellular transition. Small molecules are ranked based on cosine distance, where the molecules with the smallest one are supposed to have the highest potential to induce desired conversion. Massive optimization took place to tune the coefficients in the formula of influence score as a part of our approach. We calculated the number of experimentally validated small molecules inducing cellular conversions in the top of a ranged list and used it as a quality metric. Finally, a list of probable small molecules was obtained for various conversions.

УДК 004.94, 577.2

Веб-сервис WebMCOT для определения совместно встречаемых ДНК мотивов в данных CHiP-Seq

А.М. Мухин, В.Г. Левицкий, С.А. Лашин

Институт Цитологии и Генетики СО РАН, пр. Академика Лаврентьева 10, Новосибирск, Россия

mukhin@bionet.nsc.ru; acamukhin@gmail.com

Регуляция экспрессии генов эукариот контролируются транскрипционными факторами (ТФ), а их сайты посадки называются сайтами связывания транскрипционных факторов (ССТФ). Наиболее консервативная часть ССТФ представляется мотивом ДНК. Механизм действия ТФ, как правило, кооперативный, поэтому часто рассматривается набор двух ССТФ в составе композиционного элемента (КЭ). Мотивы в составе КЭ могут перекрываться или быть разделены спейсером. Метод MCOT, разработанный ранее в Институте Цитологии и Генетики СО РАН, может предсказывать КЭ как со спейсером, так и с перекрывающимися мотивами, используя результат одного эксперимента ChIP-Seq. Ранее такое предсказание было невозможно.

Целью данной работы является разработка веб-интерфейса WebMCOT (https://webmcot.sysbio.cytogen.ru) для кода ядра MCOT. Работа была выполнена с использованием языков программирования Python (код для сервера с использованием библиотеки Flask) и JavaScript (для реализации одностраничного веб-приложения с использованием библиотеки Vue.JS) с использованием брокера сообщений Redis. Для визуализации распределения КЭ разных структурных типов была использована библиотека Matplotlib, для визуализации тепловой карты консервативностей мотивов - скрипты на языке R. Профилировщик jeprof был использован для оптимизации кода ядра MCOT.

УДК 579.25

Risk of mitochondrial deletions is affected by the global secondary structure of the mitochondrial genome

V. Shamanskiy[1,2], A. A. Mikhailova[3,1], K. Ushakova[1], A. G. Mikhailova[1,4], S. Oreshkov[1], D. Knorre[5,6], E. O. Tretiakov[7], M. Zazhytska[8], S. W. Lukowski[9], C. Liou[10], T. Lin[10], W. S. Kunz[11,12], A. Reymond[13], I. Mazunin[14,2], G. A. Bazykin[14,15], K. Gunbin[1,16], J. Fellay[17], M. Tanaka[18,19,20], K. Khrapko[21], K. Popadin[1,13,17,22]

[1] Center for Mitochondrial Functional Genomics, Immanuel Kant Baltic Federal University, Alexander Nevsky st. 14, 236041, Kaliningrad, Russia

[2] Department of Genomic Medicine, Fomin Women's Health Clinic, Dolgorukovskaya st. 17-1, 127006, Moscow, Russia

[3] Institute for Evolution and Biodiversity, University of Münster, Hüfferstraße 1, 48149, Münster, Germany

[4] Vavilov Institute of General Genetics RAS, Gubkina st. 3, 117971, Moscow, Russia

[5] Laboratory of Systems Biology and Computational Genetics, Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Kolmogorova st. 1, 119991, Moscow, Russia

[6] Institute of Molecular Medicine, Sechenov First Moscow State Medical University, Bolshaya Pirogovskaya st. 2, 119435, Moscow, Russia

[7] Department of Molecular Neurosciences, Center for Brain Research, Medical University of Vienna, Spitalgasse 4, 1090 Wien, Vienna, Austria

[8] Department of Biochemistry and Molecular Biophysics, Mortimer B. Zuckerman Mind Brain and Behavior Institute, Columbia University, 3227 Broadway, NY 10027, New York, United States

[9] Institute for Molecular Bioscience, University of Queensland, 306 Carmody Rd, St Lucia QLD 4072, Brisbane, Australia

[10] Neurology, Kaohsiung Chang-Gung Memorial Department of Hospital and Chang-Gung University, No. 123號, Dapi Road, Niaosong District, Kaohsiung, Taiwan

[11] Division of Neurochemistry, Department of Experimental Epileptology and Cognition Research, University of Bonn, Regina-Pacis-Weg 3, 53113 Bonn, Germany

[12] Department of Epileptology, University Hospital of Bonn, Venusberg-Campus 1, 53127 Bonn, Germany

[13] Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland

[14] Center of Life Sciences, Skolkovo Institute of Science and Technology, Bolshoy blvd. 30, 143026, Skolkovo, Russia

[15] Sector of Molecular Evolution, Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences, Bolshoi Karetny per. 19, 127994, Moscow, Russia

[16] Institute of Molecular and Cellular Biology SB RAS, Academician Lavrent'ev Ave. 8/2, 630090, Novosibirsk, Russia

[17] School of Life Sciences, Ecole Polytechnique Federale de Lausanne, Route Cantonale, 1015 Lausanne, Switzerland

[18] Department for Health and Longevity Research, National Institutes of Biomedical Innovation, Health and Nutrition, Tokyo, Japan

[19] *Department of Neurology, Juntendo University Graduate School of Medicine, Tokyo, Japan*
[20] *Department of Clinical Laboratory, IMS Miyoshi General Hospital, 974-3 Fujikubo, Miyoshi, Iruma District, Saitama 354-0041, Japan*
[21] *Biology Department, Northeastern University, 360 Huntington Ave, Boston, MA 02115, United States*
[22] *Department of Life Sciences, Swiss Institute of Bioinformatics, Amphipôle, Quartier UNIL-Sorge, 1015 Lausanne, Switzerland*

v.a.shamanskiy@gmail.com

Ageing is often associated with clonal expansion of somatic mitochondrial deletions, while their origin is still poorly known. Deletions are often flanked by direct nucleotide repeats, however, repeats solely do not provide an exhaustive explanation of deletion distribution. Here, we hypothesized that repeats have higher chances to be realized into deletions in case of their spatial proximity. Analyzing the distribution of human deletions we observed a hot spot (6-9kb and 13-16kb), which is not explained by direct repeats and might be driven by close contacts of these two regions during mtDNA replication. Using several in silico approaches we reconstructed the secondary structure of the major arc and proposed that it is organized as a large-scale hairpin-like loop with a center close to 11 kb and stem between 6-9 kb and 13-16 kb. mtDNA Hi-C data of healthy and COVID-19 patient samples also demonstrated a high-density region in the expected contact zone. In our final model, we demonstrated that repeats within the contact zone are 3-times more mutagenic as compared to repeats outside the contact zone, which clarifies also well known increased mutagenicity of the common repeat (8470-8482 bp and 13447-13459 bp). The proposed topological model improves our understanding of the mechanisms of deletion formation in the human mitochondrial genome and opens a possibility to predict deletion burden in different human haplogroups and mammalian species.

УДК 577.2

HLA variants differ in binding preferences of self-peptides from proteins with specific molecular functions

V. Karnaukhov[1], W. Paes[2], I. Woodhouse[3], T. Partridge[2], A. Nicastri[4], D. Shcherbinin[5], D. Chudakov[1,5], I. Zvyagin[5], N. Ternette[4], H. Koohy[3], P. Borrow[2], M. Shugay[5]

[1] Center of Life Sciences, Skolkovo Institute of Science and Technology, Bolshoy blvd. 30, 143026, Skolkovo, Russia
[2] Nuffield Department of Clinical Medicine, University of Oxford, Oxford OX1 2JD, UK
[3] Medical Research Council (MRC) Human Immunology Unit, MRC Weatherall Institute of Molecular Medicine (WIMM), John Radcliffe Hospital, University of Oxford, Oxford, UK; MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX1 2JD, UK
[4] The Jenner Institute, Nuffield Department of Medicine, University of Oxford, Oxford OX1 2JD, UK
[5] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, Moscow, Russia; Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Pirogov Russian National Research Medical University, Ostrovityanova st. 1, 117997, Moscow, Russia

vadim-karnaukhov@mail.ru

Presentation of foreign and self peptides to T cells by human leukocyte antigen (HLA) has a key role in adaptive immune responses. Interactions between peptide and HLA binding groove residues shape HLA ligandome favoring binding of peptides having a certain sequence motif. HLA genes are highly polymorphic and the polymorphisms in peptide-contacting sites result in distinct sets of peptides presented by different HLA variants. This may lead to enrichment or depletion of HLA ligands for some proteins, and such preferences may differ between HLA variants. In this study, we investigated peptide-binding preferences of HLA in terms of functions of the presented proteins by statistical analysis of the *in silico* predicted HLA ligandomes. Our results demonstrate that HLA have a tendency to present peptides derived from proteins with specific molecular functions and these preferences are similar between the alleles with similar anchor residue preferences.This may be explained by preferential HLA presentation of the proteins enriched by the amino acids that are favourable anchor residues for that allele. Our observations can be extrapolated to explain the protective effect of certain HLA alleles in infectious diseases.We hypothesize that it can also explain susceptibility to certain autoimmune diseases and cancers. We demonstrate that these differences lead to differential presentation of HIV, Ebola, influenza virus, SARS-CoV-1 and SARS-CoV-2 proteins by various HLA alleles. Finally, we show that the reported HLA presentation bias may be compensated for in haplotypes to increase the size of the immunopeptidome presented in each individual.

УДК 004.94, 577.2

TCRen: a knowledge-based pairwise potential that accurately predicts TCR-peptide recognition

V. Karnaukhov[1], D. Shcherbinin[2,3], A. Chugunov[2], I. Zvyagin[2,3], R. Efremov[2], D. Chudakov[1,2,3], M. Shugay[1,2,3]

[1] Center of Life Sciences, Skolkovo Institute of Science and Technology, Bolshoy blvd. 30, 143026, Skolkovo, Russia
[2] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Miklukho-Maclay st. 16/10, 117997, Moscow, Russia
[3] Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Pirogov Russian National Research Medical University, Ostrovityanova st. 1, 117997, Moscow, Russia

vadim-karnaukhov@mail.ru

T cell receptor (TCR) recognition of foregn peptides presented by major histocompatibility complex (MHC) proteins is a crucial step in triggering adaptive immune response. Prediction of TCR:peptide recognition is important for many clinically relevant problems: prediction of cross-reactivity of TCRs used in adoptive T-cell based therapies, identification of targets for antigen-specific therapies of autoimmune disorders, vaccine design.

In this work we propose a knowledge-based potential TCRen that can be used to assess binding probability between TCRs and cognate antigens. TCRen is derived from statistics of amino acid residue contacts between peptides and TCRs in crystal structures of TCR-peptide-MHC complexes from PDB. We demonstrate excellent performance of TCRen for two tasks related to TCR-peptide recognition: 1) descrimination between real and mocked TCR-peptide-MHC complexes; 2) discrimination between cognate epitope and unrelated peptides in TCR-peptide-MHC crystal structures. Comparison of TCRen with potentials describing general protein-protein interaction and protein folding rules reveals the distinctive features of TCR-peptide interactions, such as intrinsic asymmetry of the interface, complex interplay between different physico-chemical properties of contacting residues and lower impact of hydrophobic interactions. Finally, we propose a computational pipeline utilizing TCRen for identification of cancer neo-epitopes recognized by tumor-infiltrating lymphocytes.

УДК 004.94, 577.2

Триплетный состав общих генов митохондрий и хлоропластов растений выявляет их дифференциацию

В. Федотовская[1], М. Садовский[1,2]

[1] Сибирский федеральный университет, Институт фундаментальной биологии и биотехнологии, Свободный пр., 79/4, 660041, Красноярск, Россия
[2] Институт компьютерного моделирования СО РАН, ул. Академгородок 50/44, 660036, Красноярск, Россия

minaeva.mav@gmail.com

Выявление связей и закономерностей, связывающее структуру нуклеотидных последовательностей, их функцию и таксономию их носителей — одна из фундаментальных задач биоинформатики.

Эта связь изучалась на примере генов ATP-синтазы митохондрий и хлоропластов растений. При этом изучались растения, для которых депонированы гены обеих органелл одновременно. Было исследовано 85 таких растений. Гены извлекались из полногеномных последовательностей с помощью программы CLC Genomics Workbench (*atp1, atp4, atp6, atp8, atp9* из митохондриального генома и *atpA, atpB, atpE, atpF, atpH, atpI* из хлоропластного). Каждая последовательность генов преобразовывалась в частотный словарь триплетов. Частотный словарь триплетов — это информация о частоте встречаемости всех триплетов, определяемой по последовательности генов. При этом рамка считывания двигалась с шагом $t = 1$. После этого проводилась кластеризация методом упругих карт с помощью ПО ViDaExpert.

Исследовались следующие ключевые вопросы:

1. выделяются ли кластеры на упругой карте?
2. если да, то каким образом состав кластеров связан с принадлежностью точек к органеллам, функции генов и таксономии носителей?

Для всех типов кластеризации было обнаружено четкое разделение генов, кодирующих одну и ту же субъединицу АТФ-синтазы, в отдельные кластеры. Таким образом было доказано преобладание функции над таксономией для семейства генов АТФ-синтазы геномов митохондрий и хлоропластов растений. Также было обнаружено 3 гена, выбивающихся из общего распределения: *atp4* (*Welwitshia mirabilis*), *atp6* (*Ammopiptanthus mongolicus*), *atpF* (*Ammopiptanthus nanus*).

УДК 575.8, 579.25

Филогенетическая классификация и биосинтетический потенциал штамма *Rhodococcus* S11

А. Грицева, М. Маркелова, И. Хиляс

Казанский Федеральный Университет, ул. Кремлевская, 18, 420008, Казань, Россия

gritseva42@gmail.com

Представители рода *Rhodococcus* отличаются широким метаболическим разнообразием и устойчивостью к экстремальным условиям окружающей среды. Несмотря на их перспективность и постоянные изменения таксономической классификации, важностью разрешения филогенетических отношений между представителями рода *Rhodococcus* часто пренебрегают.

В данной работе был произведен комплексный геномный и таксономический анализ нового штамма *Rhodococcus fascians* S11, изолированного из засушливых серпентинитовых пород Халиловского массива в Оренбургской области. Для описания метаболического профиля проводился поиск кластеров генов, вовлеченных в биосинтез вторичных метаболитов, и фенотипический анализ на системе для идентификации и исследования свойств микробных культур от BiOLOG с использованием тест-панели GEN III MicroPlate.

Сравнение последовательностей полных геномов и выбранных маркерных генов (16S рРНК, *gyr*B, *cat*A) штамма S11 с другими представителями рода *Rhodococcus* показало, что S11 принадлежит к виду *R. fascians*. При анализе пангенома был обнаружен высокий уровень геномной изменчивости и пластичности внутри вида. Были выявлены кластеры NRPS и PKS, ответственные за биосинтез вторичных метаболитов, способствующих адаптации к экстремальным условиям, а также кластер генов эктоина – эффективного осмолита. Функциональная аннотация RAST и фенотипический анализ свидетельствовали о наличии у штамма S11 адаптивных особенностей, обусловленных избыточностью генов, ответственных за стрессовые реакции, устойчивостью ко многим соединениям и способностью расти при высоких концентрациях солей (NaCl до 8%).

Работа поддержана стипендией Президента Российской Федерации для молодых ученых и аспирантов.

ИНСТИТУТ БИОИНФОРМАТИКИ
СБОРНИК ТЕЗИСОВ 2020/21

Подписано в печать 06.07.2021. Формат 60×84/16. Печать цифровая.
Усл. печ. л. 5,5. Тираж 36. Заказ 3186.

Отпечатано с готового оригинал-макета, предоставленного
ООО «Институт биоинформатики»
в Издательско-полиграфическом центре Политехнического университета.
195251, Санкт-Петербург, Политехническая ул., 29.
Тел.: (812) 552-77-17; 550-40-14.