BIOINFORMATICS
INSTITUTE

RESEARCH PROJECTS
BIOINFORMATICS INSTITUTE

2018/2019

BIOINFORMATICS INSTITUTE 2018/19

Project abstracts

Saint Petersburg

2019

BIOINFORMATICS INSTITUTE 2018/19.
Project abstracts
Saint Petersburg, 2019.

Table of contents

FALL 2018.....5

Analysis of factors affecting the course of chronic myeloid leukemia.....5

Detection of interchromosomal rearrangements from Hi-C data.....6

A library of functions for express analysis of FASTA/FASTQ files.....7

Visualization of signaling pathways basing on genes differential expression profile.....9

The comparative analysis of MDR *Klebsiella pneumoniae* genome.....10

Comparative analysis of the human pathogens genomes *Neisseria meningitidis*.....12

Comparative analysis of NUMT in underground and terrestrial rodents…14

The study of processes of gene gain and loss within *Lactobacillus* species…15

Modeling of mouse chromosome banding pattern.....16

Prioritization of genetic variants.....17

Assembly of yeast genome with Oxford Nanopore data.....18

The effect of X chromosome inactivation on the expression of autosomal genes.....19

Finding of cis-regulatory elements in promoters.....20

Study spectrum of genetic variants in TTN gene.....21

Transcriptional response of pea roots to symbiosis markers.....22

De novo assembly and analysis of *Platynereis dumiliii* (Nereididae, Annelida) transcriptome at different stages of regeneration.....23

Plasmid host range prediction based on CRISPR arrays. Plasmids CRISPR Cas systems search.....24

Analysis of nonsense alleles of *Caenorhabditis elegans* genes.....26

Increasing the length of introns due to transposable elements.....27

Analysis of the Drosophila melanogaster full genome sequences.....28

Automated pathway annotation for single-cell RNA-seq.....29

Automated marker descriptor for single-cell RNA-seq.....30

SPRING 2019.....31

Antimicrobial peptide prediction in non-model species based on transcriptome data.....31

***Denoising of ULI-NChIP-seq data with neural networks*.....32**

Noisy peak calling.....33
Identification of pathway genes triggered differential expression profile changes.....34
Improving peak calling in SPAN.....35
Detection of pathogenic INDELs and SNPs in whole-exome sequencing data of patients with different types of idiopathic cardiomyopathy.....36
Diversity of opsins in transcriptomes of Baikal endemic amphipodes.....37
Whole-genome Drosophila sequence analysis - 2.....38
The influence of molecular dynamics parameters on protein motion characteristic timescale.....39
Association rule mining using fishbone diagrams.....40
Analysis of yeast genomes from the Peterhof genetic collection.....41
Towards detection of differential RNA editing events in transcriptomics datasets.....42
Searching for latent viruses in human whole genome sequencing data…….43
Adaptation of fish to the depth.....44
Local sequence alignment using intra-processor parallelism.....45
Effect of smoking on human leukocyte epigenome.....46
Detection of CNVs in patients with different types of idiopathic cardiomyopathies.....48
Bayesian optimization for demographic history inference.....49
Implied weighting as a measure of clade support: automation of the task and comparative assessment of results.....50
Systematic comparison of state-of-the-art variant callers' performance.....51
SPAdes support for third-party assembly graphs.....52
SUMMER 2019.....53
Identification of pathway genes triggered differential expression profile changes.....53
Improving peak calling in SPAN.....55
Searching of potential markers of pregnancy complications using sequencing data of circulating cfDNA from maternal plasma.....56
Bayesian optimization for demographic history inference.....58

FALL 2018

Analysis of factors affecting the course of chronic myeloid leukemia

I.Babkina[1] , N.Pogodina[1] , O. Stanevich[2] , E.Bakin [3]

[1]Saint Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
e-mail: kriska.irichka@gmail.com, nadezda.pogodina.bio@gmail.com
[2]Pavlov First Saint Petersburg State Medical University
6-8 L'va Tolstogo str., Saint Petersburg, 197022, Russia
e-mail: oksana_stanevich@gmail.com
[3] Bioinformatics institute
2A Kantemirovskaya, Saint Petersburg, 197342, Russia
e-mail: eugene.bakin@gmail.com

Chronic myeloid leukemia (CML) is a myeloproliferative disorder characterized by unregulated granulocytic proliferation. A standard treatment includes tyrosine kinase inhibitor (TKI) and hematopoietic stem cell transplantation (HSCT). In this project, our goal was to identify the factors that have an impact on survival after HSCT.

The most common method of survival analysis is a Kaplan-Meier approach. These method works on censored data, when observation ended up before event of interest occurred. We plotted overall and event-free survival curves. Cumulative survival probability was 37.9%. Then, we performed analysis of single variables (conditioning regimens, phase of CML et al.) and compared survival curves in log-rank test. We identified one statistically significant factor: cyclophosphamide therapy after HSCT (p-value = 0.03).

CML therapy has been improved, so we used multivariate analysis to assess the influence of new treatment methods on the survival using correlation test. Correlation matrix showed weak association, that's why we selected the following factors with the specialist's help: conditioning regimens, phase of CML, graft compatibility, TKI therapy, cyclophosphamide therapy after HSCT. Ordination methods (PCA and MDS) showed 3 clusters of therapy factors, distributed between 3 era: 1995-2006, 2007-2012, 2013-2018. The eras' survival curves were also statistically different (p-value = 0.009).

Detection of interchromosomal rearrangements from Hi-C data

N. Alexeev[1], D. Orekhov[2], E. Kartysheva[2]

[1] ITMO University
49 Kronverksky Pr., Saint-Petersburg, 197101, Russia
[2]Saint Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia

Chromosomal rearrangements disturb complex 3D structure of eukaryotic genome and may lead to various disease among which is cancer, detecting them may be useful in early diagnostics. Hi-C is a relatively recent sequencing method that estimates 3D proximity between different regions of a sequenced genome, this type of data allows for detection of different chromosomal abnormalities.

We have developed an algorithm that scans through Hi-C map and reports the presence of interchromosomal rearrangements with the coordinates of their breakpoints. The algorithm relies on 2D convolution and GMM for filtering out the data and detection of interchromosomal interactions, then a sliding-window approach is used for breakpoint localization. The method is tested on Hi-C maps obtained from glioblastoma cells of *H.Sapiens*, showing both high precision and high recall.

A library of functions for express analysis of FASTA/FASTQ files

A. Kizenko[1], A. Morshneva[1], P. Pavlova[1], E. BakinBakin[2]

[1]Saint-Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
e-mail: alyona.kizenko@gmail.com
[2]Bioinformatics Institute
2A Kantemirovskaya, Saint Petersburg, 197342, Russia
e-mail: eugene.bakin@gmail.com

Frequently, when carrying out bioinformatics projects including FASTA/FASTQ files processing, one has to solve routine tasks, e.g. deduplication of sequences. A common approach for this is writing little scripts in Python/bash or dealing with existing programs, which may be complicated for usage. Therefore, we decided to create a flexible tool containing functions for processing files with sequencing data.

We created a program called BreakFAST which is based on Python 3 and the following libraries: Biopython, argparse, pandas, numpy, matplotlib and re. The tool consists of three modules.

Basic statistics module can be used for counting:

- minimum, maximum, mean, total length of reads;
- GC-content;
- quality scores;
- N bases.

Filtering module can be used for deleting:

- reads shorter than X;
- reads containing Ns;
- poor quality reads;
- duplicates;
- reads with a particular motif.

Matching module can be used for:

- joining reads from files;
- finding overlapping between files;
- subtracting sets of reads from files.

While applying commonly suggested Biopython functions we've faced performance problems while parsing a large volume of data. For mitigation of

this effect while iterating over FASTA/FASTQ files, we compared SeqIO.parser and Iterator from Biopython.

We have found that usage of Iterator in Filtering and Matching modules was optimal for iteration (10 times speed gain). Notably, we compared function “delete reads shorter than X” with the same Trimmomatic’s function and found that BreakFAST occupies up to 7 times less RAM, which may be useful when a computer’s capacity is limited. As a result, BreakFAST is a simple and customizable tool, which can be potentially updated with new modules and functions.

Visualization of signaling pathways basing on genes differential expression profile

S.V. Legkovoy[1] , O.V. Romanova[2], E.V. Bakin[3], O.V. Stanevich[4]

[1] Saint-Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
[2]The St.Petersburg State Health Care Establishment the City Hospital №40,
9 Borisova str., Sestroretsk, 197706, Russia
[3]Bioinformatics institute
2A Kantemirovskaya, Saint Petersburg, 197342, Russia
[4]First Pavlov State University of St. Petersburg
6-8 Leo Tolstoy street, Saint Petersburg, 197022, Russia

In recent years, Gene Expression Omnibus (GEO) NCBI database has accumulated a significant amount of data obtained via mRNA microarrays, which are widely used for ananalysis of differential expression profile. During a research of genes expressions, a proper visualization of results is an important task. One of the best ways is to use R language and related packages for statistical analysis, preprocessing and visualization of expression data through interaction with KEGG database.

An aim of this study was to visualize signaling pathways according to the gene expression data obtained from GEO NCBI database. To achieve our goal, we implemented an easy-to-use script based on pathview, gage and GEOquery packages, which allowed us to obtain gene expression data directly from the GEO NCBI database and to find the most significant signaling pathways from KEGG PATHWAY.

Using the developed script, we analyzed the Affymetrix microarray data, identified and visualized the most significant signaling pathways involved in reprogramming of lymphatic endothelial cells infected by human Kaposi’s sarcoma-associated herpesvirus (KSHV).

The comparative analysis of MDR *Klebsiella pneumoniae* genome

A. Kapanina[1], N. Lukashina[2], D.V. Likholetova[3], E. Bakin[4],O. Stanevich[5], S.S. Sidorenko[6]

[1]*Jetbrains limited liability company, Universiteskaya emb, 7-9-11, 5A, Saint-Petersburg, 199034, Russia*

[2]*Peter the Great St. Petersburg Polytechnic University, Politekhnicheskaya str., 29, Saint-Petersburg, 195251, Russia*

[3]*Saint-Petersburg State University, Saint-Petersburg, Universiteskaya emb., 7-9, 199034, Russia*

[4]*Bioinformatics institute,Kantemirovskaya, 2A , Saint-Petersburg, 197342, Russia*

[5]*Pavlov First Saint Petersburg State Medical University, 6-8 L'va Tolstogo str., Saint Petersburg , 197022, Russia*

[6]*Children's Scientific and Clinical Center for Infectious Diseases, Saint-Petersburg, Professora Popova street, 9, 197022, Russia*

Klebsiella pneumoniae is a gram-negative bacteria that is known as opportunistic, hypervirulent, and multidrug resistant hospital pathogen. The problem of resistance to carbenemase group of antibiotics makes it one of the main threats during hospitalisation. The diversity of *K. pneumoniae* is studied by whole-genome sequencing (WGS) and multiple typing methods including multi-locus sequence typing (MLST), that separate strains into different lineages.

In our study we assembled and analysed genomes of 22 isolates of different years and sources from the Saint-Petersburg hospitals to identify their origin and describe their pangenome.

With use of Kleborate tool, we found that our strains belong to common european and asian MLST types (ST147, ST11, ST340 and ST395). All of them carry NDM-1 and ParC resistance genes, and only one - OXA-48. According to the genes discovered in strains, we listed inefficient antibiotics for their treatment.

Via PlasmidFinder we detected a presence of plasmid R27 of *Salmonella typhi*, that can be explained by contamination of samples or by horizontal transfer between *K.pneumoniae* and *S. typhi*.

According to an existing literature, the obtained MLST types are spread in Europe and Asia. However, for obtaining a more detailed result about an origin of the strains, a genome structure analysis is needed.

In conclusion, we can say that within the period from 2012 to 2016 there were no invasions of new sequence types on a territory of mentioned hospitals.

The obtained results of pangenome analysis can be used in treatment prescription.

Comparative analysis of the human pathogens genomes *Neisseria meningitidis*

A. Matiiv[1], I. Sheshukov[1], E. Bakin[2], O. Stanevich[3], S. Sidorenko[4]

[1]*Saint Petersburg State University*
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
e-mail: antonmatiiv@yandex.ru, sheshukov.ilya@gmail.com
[2]*Bioinformatics institute*
2A Kantemirovskaya, Saint Petersburg, 197342, Russia,
e-mail: eugene.bakin@gmail.com
[3]*Pavlov First Saint Petersburg State Medical University*
6-8 L'va Tolstogo str., Saint Petersburg, 197022, Russia
e-mail: oksana_stanevich@gmail.com
[4]*Children's Scientific and Clinical Center for Infectious Diseases, Saint-Petersburg,*
Professora Popova street, 9, 197022, Russia,

Neisseria meningitidis or meningococcus often colonizes the mucous membrane of the oropharynx, causing no visible symptoms, but is also the main cause of bacterial meningitis and sepsis throughout the world. The epidemiological profile of *N. meningitidis* varies in different populations, and over time, the virulence of meningococcus is based on the plastic genome and the expression of certain capsular polysaccharides and non-capsular antigens. Twelve different serogroups based on the polysaccharide capsule have been identified, but only six of them (A, B, C, W, X and Y) account for 90% of the invasive meningococcal disease worldwide. Seven housekeeping genes for meningococcal strains are used for MLST (multilocus sequence typing) to determine their sequence types (ST).

The aim of our work was to compare whole genome sequencing data of 20 *Neisseria meningitis* samples isolated from carriers and sick people, to use phylogenetic analysis and to find a connection with antibiotic resistance, virulence and carriage.

Before the analysis of sequences, we have written a computer script for interfacing and downloading reference genomes from NCBI. We analyzed the antigen-encoding, virulence and carriage associated and antibiotic resistant gene profiles. We also searched for amino acid changes leading to penicillin resistance. To estimate the relationship between samples, phylogenetic trees were constructed on the basis of isolates assemblies by using CSI Phylogeny and

REALPHY. We constructed phylogenetic trees for carriage associated genes to figure out if the samples would cluster according to their origin of isolation.

Comparative analysis of NUMT in underground and terrestrial rodents

E. Sytnik[1], O. Bondareva[2]

[1]*Peter the Great St. Petersburg Polytechnic University*
Politekhnicheskaya str., 29, Saint-Petersburg, 195251, Russia
[2]*Zoological Institute RAS, St. Petersburg, 199034, Russia*

NUMT (nuclear mitochondrial DNA segment) is a transposition of mitochondrial DNA into nuclear genome. They are found in all eukaryotes but significantly differ in length and number among different species. Particular factors that can be associated with NUMTs are still not determined. Due to the specificity of mitochondrial genes, habitat conditions may be one of the factors.

They aim of this study was to estimate the number of NUMT for underground and terrestrial rodents. For this work we only analyzed long (>300 n.p.) NUMTs of protein-coding regions. We used Genbank database for mitochondrial and nuclear genomes (4 species for each group) and BLAST for NUMT searching. It was found that some genes like ND4L and ATP8 are not likely to be included in NUMTs, which may be caused by the small size of the genes. Some underground species are shown to have a larger amount of long NUMTs but it is yet unclear if the same is true for whole group.

Further study should include larger amount of species and dN/dS analysis for each gene to determine whether some of the NUMTs may have a functional role.

The study of processes of gene gain and loss within *Lactobacillus* species

A. Kosolapova[1,2], O. Bondareva[3]

[1]*All-Russia Research Institute for Agricultural Microbiology Pushkin, St. Petersburg, 196608, Russia*
[2]*St. Petersburg State University, St. Petersburg, 199034, Russia*
[3]*Zoological Institute RAS, St. Petersburg, 199034, Russia*

Lactobacillus genus includes Gram-positive non-sporulating bacteria known for their ability to produce lactic acid as a result of carbohydrate fermentation. To date more than 180 species refer to *Lactobacillus* genus. A hallmark of that genus is a high level of intra-group diversity. Firstly, the diversity exhibits in ecology of the group as lots of *Lactobacillus* species are associated with cavities of human and animals, for example gastrointestinal tract and urogenital tract, while others can be found on plants, in dairy and fermented products. Secondly, the genome size of *Lactobacillus* bacteria can vary between 1.2 Mb and 5 Mb. The aim of this work was to study connection between ecological specificity and genome organization within various strains of *Lactobacillus* and analyze influence of ecological specificity on processes of gain and loss of genes.

As a data for analysis we used protein and CDS sequences for 185 *Lactobacillus* species (1708 strains) from RefSeq database. *Lactococcus lactis* subsp. *lactis* Il1403 protein and CDS sequences were used as an outgroup. We classified species into 7 groups based on ecological niche. We revealed orthologous proteins within strains using Proteinortho5/POFF software. Further research should involve a phylogenetic tree reconstruction based on full orthologous genes groups followed by gain-loss analysis performed with GLOOME software.

Modeling of mouse chromosome banding pattern.

Y. Lebeda[1], Y. Barbitoff[2]

[1]*Pavlov First Saint Petersburg State Medical University*
6-8 L`va Tolstogo str., Saint Petersburg, 197022, Russia
e-mail: 1.u.r.o.n.3@gmail.com
[2]*Bioinformatics institute*
2A Kantemirovskaya, Saint Petersburg, 197342, Russia
e-mail: barbitoff@bk.ru

Differential chromosome staining is a method of chromosome staining with special dyes to detect certain discs or regions of the chromosome (also called chromosome bands). The resulting banding pattern is an important marker of genome architecture; however, no specific molecular determinants of it are known to date. Previously, our group discovered the relationship between the pattern of differential staining of chromosomes and several genomic features (ChIP-Seq tracks of Smc1a/Smc3, CTCF, polyA and polyT repeats). However, when validation of this relationship using the genome of *Mus musculus* was attempted, it was found that the distribution of genomic elements within the *M. musculus* bands differs from that observed in humans. In this project, we took an effort to develop a model that would be able to predict the border regions between bands, on the basis of the human genome data, and apply this model to predict the banding of *M. musculus* chromosomes.

We built a random forest model to predict the borders of the bands based on the number of genomic elements (i.e., ChIP-Seq peaks or k-mers) lying in the intervals of a given width inside and outside of the band borders. For prediction, the *M. musculus* genome was cut into intervals of the same width by a sliding window; and the resulting intervals were annotated with the same features that were used to train the model. Unfortunately, all the models constructed (despite high cross-validation AUC scores) failed to provide reasonable predictions – both for the mouse and human genomes, the results of the prediction of band boundaries differed from the already existing markup. The results can be explained by a large number of false-positive results, which becomes significant even with a small false-positive rate at large numbers of trials. Hence, a new model has to be sought for to explain the nature of chromosome bands.

Prioritization of genetic variants

Vasiliy Isaev[1,2], Liubov Lonishin[1,2],Yury Barbitoff[2]

[1]Peter the Great St. Petersburg Polytechnic University
Politekhnicheskaya str., 29, Saint-Petersburg, 195251, Russia
[2]Bioinformatics institute
2A Kantemirovskaya, Saint Petersburg, 197342, Russia

The identification of deleterious mutations within candidate genes is a crucial step in the elucidation of the genetic bases of human disease, consequently, there is a need to aim attention at classifying appropriate mutations. The goal of our programme, which is called MutationsPriorityPredictionTool (MPPT) is to find out these genetic variants from thousands of others in order to help clinicians and geneticists.

To calculate the coefficient we have developed an tool (https://github.com/Vasiliy566/Mutations-Priority-Prediction-ToolVCFparser/). Tool get on input vcf-file and set of simple configurations that contain rules on how to calculate mutation priority score depends from parameters given in the file. After calculating tool will print top of mutation by parameter specified by user. It can be top 10% of mutations or 100 mutations ot other option.

We have tested our programme on whole exome sequencing data, obtained from the resource centre. First of all, the selection of the test sample was made in accordance with the ClinVar database and was compared with results of Franklin (https://franklin.genoox.com/), which is based on ACMG recommendations (The American College of Medical Genetics and Genomics). The percentage number of correct calls by MPPT was calculated, and the sensitivity and specificity of the method was determined.

Accuracy of our programme is 67,5%, sensitivity is about 100% (95% CI = 79.4% to 100.00) and specificity is 60,6% (95% CI =53.9% to 67.3%). Testing on the whole data, we obtained 114 mutations above the threshold from more than 22 thousand at all.

MPPT focuses on pathogenic variants without losing them, but also keeps some benign variants which should be manually checked by a specialist after running. In the future, we will add this functionality to NGB (New Genome Browser).

Assembly of yeast genome with Oxford Nanopore data

Andrew G. Matveenko[1,2], Yury A. Barbitoff[1,2], Alexander V. Predeus[1,3]

[1]Bioinformatics institute
2A Kantemirovskaya, Saint Petersburg, 197342, Russia
[2]Saint Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
[3]University of Liverpool, UK, Liverpool, L69 3BX

Baker's yeast *Saccharomyces cerevisiae* is a widely used model organism. The Peterhof genetic collection (PGC) is a large laboratory stock unrelated to the yeast reference strain. Previously, several PGC strains were sequenced using Ion Torrent technology. However, the resulting assemblies were incomplete and required substantial improvement. We attempted to obtain a reference quality assembly of one PGC strain, 1A-D1628, using Oxford Nanopore Technology (ONT) sequencing.

Raw data was obtained from one ONT MinIon flowcell which generated 10.15 Gbp total sequence length (836x coverage). To create draft genome assembly we used three long-read assemblers: Canu, Flye and wtdbg2. Canu produced the best results, with 17 large (> 50 kbp) contigs that correspond to 16 yeast chromosomes and mitochondrial DNA. Flye was slightly worse with 18 large contigs as it failed to assemble chromosome III as a single molecule. Wtdbg2 failed to produce any sensible sequence. Comparison of the Canu assembly with the reference showed that it contained 105 misassemblies, and large amount of mismatches and short indels. We also analyzed structural variations in the strain using NGMLR-Sniffles pipeline. The results of analysis were concordant with variations described previously in 1A-D1628.

In conclusion, the data obtained from the Oxford Nanopore sequencing can be used to analyse structural variations in the 1A-D1628 strain. However, the de novo assembly requires additional correction and polishing to reach the reference quality. Several strategies can be used to achive this goal. First, an alternative basecaller can be used to improve the quality of reads. Second, exclusively for ONT, the assembly can be improved by polishing with the MinIon raw signal using Nanopolish tool. And finally, polishing the assembly with the obtained Illumina reads should improve the accuracy of the sequence producing high quality reference, which can be used for comparative genomic studies.

The work is supported the RSF grant 18-14-00050.

The effect of X chromosome inactivation on the expression of autosomal genes

D. Kilina[1], Y.Barbitov[2], R.Skitchenko[2]

[1] Almazov National Medical Research Centre
2 Akuratova Street, Saint Petersburg, 197341, Russia
e-mail: dasha-kilina@mail.ru
2 Bioinformatic institute
2A Kantemirovskaya Street, Saint Petersburg, 197342, Russia
e-mail: barbitoff@bk.ru

X chromosome inactivation (XCI) silences the transcription of genes located on one of the X chromosomes to balance expression dosage between XX females and XY males. According to a recent work by Tukiainen et al., there is no total X chromosome inactivation in humans as up to one-third of X-chromosomal genes are expressed from both the active and inactive X chromosomes in female cells. However, the effects of XCI on the expression profile of autosomal genes have not yet been assessed. In this study we compared the expression of autosomal genes in cells with different active X chromosome copies.

To answer this qestion we analyzed public experimental data of single-cell RNA-sequencing of pancreatic islets from one female individual. We aligned the reads coming from each cell to a reference genome assembly using bowtie2. We then performed variant calling with samtools/bcftools in order to group the cells by active X chromosome by visual inspection of the alignments and SNP calls in IGV. We grouped the cells by alleles at variant sites in the XIST gene that is totally expressed from only one X chromosome. We then quantified gene expression levels with RSEM and used LIMMA plugin in the Phantasus browser to compare gene expression in the two groups of cells defined above.

We identified some candidate genes, expression of which depends on the active X chromosome copy; however, the difference in the expression levels of these genes between groups was not significant (P-value < 0.05, adjusted P-value > 0.05). Furthermore, we observed lack of clear separation between the two groups of cells based on principal component analysis (PCA), which may indicate confounding effect of cell types or other factors. Hence, the effect of differential genes expression from the X chromosomes on the expression profile of autosomal genes needs further investigation.

Finding of cis-regulatory elements in promoters

D. Balashova[1,2], E. Polyakova[2]

[1]Lomonosov Moscow State University, GSP-1
Leninskie Gory, Moscow, 119991, Russian Federation
e-mail: dashabalashova@gmail.com
2 Bioinformatics Institute
2A Kantemirovskaya street, Saint Petersburg, 194100, Russia
e-mail: enterlina@gmail.com, phone: +375(44)598-81-86

We consider a genome-wide statistical approach for the detection of specific DNA sequence motifs based on similarities between the promoters of similarly expressed genes. A comprehensive landscaping of major regulatory motifs can contribute to understanding molecular mechanisms of many complex diseases.

Assuming position-specificity of the function of promoter motifs, providing gene expression data of reasonable measurements of the number of transcripts and reflecting of the activity of the promoter, we develop *cisExpress* software that includes the algorithm for finding statistically significant associations between words of defined length with respect to the transcription start site in the expression dataset. Subsequent optimization includes combining motifs that have small differences and clustering basic words of fixed size into larger composite motifs. The analysis of time series, conducted on the basis of Hidden Markov Models, allows us to observe the significance of the found motifs over time. The tool is complemented by interactive graphical representations.

Study spectrum of genetic variants in TTN gene

O.Lebedenko[1], A.Kiselev[2]

[1]Peter the Great Saint Petersburg Polytechnic University
29 Polytechnicheskaya str., Saint-Petersburg, 195251, Russia
e-mail: oolebedenko@gmail.com
[2]Federal Almazov Medical Research Centre
Akkuratova str.,2, St.Petersburg, 197341, Russia
e-mail: artem.kiselyov@gmail.com

The TTN gene with 363 coding exons encodes titin, a giant muscle protein spanning from the Z-disk to the M-band within the sarcomere. Titin has roles in assembling and maintaining sarcomere structure, flexibility, stability, stretch and force transmission. Mutations in the TTN gene have been associated with various cardiomyopathies.

The main aim of this study was to investigate spectrum of genetic variants in TTN gene within group of patients with cardiomyopathy. 151 different type of cardiomyopathy samples, sequenced with Haloplex custom targeted capture, were processed with SNP Calling pipeline implemented on Snakemake and annotated by snpEff. Among 418 discovered SNP 64.44% variants were missens and 35.56% variants were silence. The PCA analysis showed absence of clustering SNP by type of cardiomyopathy. Fisher's exact tests with Bonferroni correction were used to compare allele frequencies of observed variants against all gnomAD population. Pathogenicity of 12 discovered statistically meaningful missense variants was predicted by algorithms SIFT, PolyPhen-2, Mutation Assessor, Provean and I-Mutant 3.0.

Almost all variants showed neutral effect on protein structure and stability. The most interesting SNP was the mutation rs9808377 I62T, presumably affecting on stability of the subunit Fn3-102 titin by I-Mutant 3.0. Presumably, this result may be explained by difficulties in multiple comparison connected with a high rate of spontaneous mutation owing to enormous size of TTN gene. Another reasons in analyzing accompanying TTN variants in cardiomyopathy group with confirmed well-knowing causative mutations.

Transcriptional response of pea roots to symbiosis markers

V.E. Tvorogova[1], P.Y. Kozyulina[2], E.A. Dolgikh[2]

[1]Saint Petersburg State University
7/9 Universitetskaya emb., Saint Petersburg, 199034, Russian Federation
[2]All-Russia Research Institute for Agricultural Microbiology
Podbelsky chausse 3, Pushkin, St. Petersburg, 196608, Russian Federation

Root nodules in legumes are symbiotic organs hosting nitrogen-fixing bacteria. At the beginning of the formation of nodule, bacteria enter the intercellular space of the root; therefore, the host plant needs an accurate recognition system that allows it to let symbiotic bacteria pass inside its tissues and block parasitic organisms from doing the same. The main external signals that provide such recognition are chitooligosaccharides of different lengths. Thus, chitooligosaccharides consisting of five monomers (co5) are markers of symbiotic bacteria, while chitooligosaccharides consisting of eight monomers (co8) are markers of parasitic organisms (insects and fungi).

The purpose of this study was to analyze the data of MACE-sequencing of pea (*Pisum sativum*) RNA from roots pretreated with co5 or co8 chitoligosaccharides.

Using the pea nodule transcriptome obtained previously (Zhukov et al., 2015) and the Dedupe software from BBTools package, we removed ambiguous transcripts and got the optimal reference transcriptome for our data. Then, using the DESeq2 and GSEABase packages, we analyzed differential gene expression in our samples and performed gene enrichment analysis. According to the results obtained, co5 treatment shows more prominent differential gene expression compared to co8 probably due to incomplete reference transcriptome. However, both co5 and co8 chitoligosaccharide treatments activate gene sets that are responsible for parasite-host interaction, chitin binding and cleavage, as well as numerous signaling pathways which include different phytohormones, receptor kinases and transcription factors.

Zhukov, V.A., Zhernakov, A.I., Kulaeva, O.A., Ershov, N.I., Borisov, A.Y., and Tikhonovich, I.A. (2015). De Novo Assembly of the Pea (Pisum sativum L.) Nodule Transcriptome. Int J Genomics 2015, 695947.

De novo assembly and analysis of *Platynereis dumiliii* (Nereididae, Annelida) transcriptome at different stages of regeneration

N.V.Zenkova[1], R.H.Abasov[2], M.A.Nesterenko[1]

[1] Saint-Petersburg State University
7/9 Universitetskaya Emb., St. Petersburg, 199034, Russia
[2]Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology,
1 Samory Mashela str., Moscow, 117997, Russia

Regeneration – the regrowth or repair of cells, tissues and organs – is widely but non-uniformly represented among all animal phyla. However, the potency of its highly variable even within a single group. The object of this study is the polychaeta *Platynereis dumerilii* (Nereididae, Annelida), capable to recover only tail. RNA-seq data of different time points after amputation (0, 4, 12, 24 hours, 2 and 4 days) from “head” and “tail” sites of regeneration were analyzed. Libraries of corrected read pairs (Karect, Trimmomatic, BBtools) were used to the de novo assembly of reference transcriptome (Trinity). The resulting assembly was characterized by high quality (TransRate-score = 0.2441) and completeness (BUSCO vs Metazoa-odb9 = 99.5%). The amino acid sequences predicted by TransDecoder (N = 160381) were compared to the Swiss-Prot database using the Diamond (e-value = 1e-10). More than 61% of the sequences were successfully annotated, but among the sequences without hits we assume the presence of species-specific proteins. Based on the normalized expression levels analysis results (Salmon), sets of “associated” sequences were highlighted for each of the samples. We suggest that incomplete overlap between “associated” sets both between time points and between sites indicate complex dynamics of gene activity during postamputation events. However, expression patterns of regeneration conservative genes (for instance: Piwi-, Vasa-, Wnt- and Notch-like) varies slightly between “head” and “tail” sites. Based on the results obtained, it can be assumed that cell proliferation is not over on 4 days after amputation and damaged structure recovery will be observed at later stages of generation.

Plasmid host range prediction based on CRISPR arrays. Plasmids CRISPR Cas systems search

M.U.Kongoev[1], I.V.Fedorova[2], M.P Rayko[3]

[1]ITMO University (Saint Petersburg National Research University of Information Technologies, Mechanics and Optics), Saint Petersburg, 197101, Russia
e-mail: mikhailkongoev@mail.ru
[2]Skoltech, Skolkovo Institute of Science and Technology, Moscow, 121205, Russia
e-mail: femtokot@gmail.com
[3]Center for Algorithmic Biotechnology St. Petersburg State University, Saint-Petersburg, 199004, Russia
e-mail: mike.rayko@gmail.com

Horizontal gene transfer plays a highly important role in evolution of bacteria. Presumably, gene exchange between bacteria occurs by genetic mobile elements such as plasmids and bacteriophages. But for nowadays there is no reliable way to check if the certain plasmid can “travel” between bacteria of different origin, and how wide the plasmid host range could be.

Also it is important to be able to predict plasmid host in case of metagenomic data, where we usually have dozens of novel plasmids without any information of host species.

To answer this question, we analyzed CRISPR cassettes in bacterial genomes – repetitive sequences in bacterial DNA, interspaced with unique “spacer” sequences, which were extracted from genetic mobile elements infected the bacteria or its ancestors. Spacers in CRISPR cassette can be considered as a link between the plasmid and its host.

We used CRISPR Finder spacers database and the RefSeq database of all plasmids known to date (November 2018). Blasting spacers over plasmids sequences allowed us to determine plasmid host ranges: variety of bacterial organisms where the plasmid can exist.

By taxonomy analysis we found some plasmids which can live in different families of organisms, they can be useful in genetic engineering as a natural shuttle vectors.

Taxonomy analysis showed that a bunch of plasmids have additional hosts except of host they were related to according to RefSeq database: 543 blast hits – additional hosts of different genus, 29 blast hits - different family, 19 - different order, 12 - different class and even 2 blast hits – additional hosts of different phylum! Thus, plasmids are actually “travelling” between bacteria species and can be important players in process of evolution.

We also found, that a lot of plasmids carry their own defense CRISPR systems (10% of RefSeq plasmids). Part of these systems (10%) seems to be active – there are Cas1 genes near CRISPR cassettes. Role of these systems in plasmid propagation, host fitness and evolutionary relationship with the known chromosomal CRISPR-Cas systems is the subjects of future research.

Analysis of nonsense alleles of *Caenorhabditis elegans* genes

D. Chaplygina[1], N. Potapova[2]

[1]*Saint Petersburg State University*
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russian Federation
e-mail: das9884@gmail.com
[2]*Lomonosov Moscow State University*
GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation
e-mail: nadezhdalpotapova@gmail.com

Nonsense mutation in gene is a mutation that results in a premature stop codon. Most of genes with nonsense alleles translates into a nonfunctional proteins, which makes such genes to be a pseudogenes.

The purpose of this study was to analyze the distribution of nonsense mutations in *Caenorhabditis elegans* genes and to perform a direct measurement of the strength of negative selection acting on nonsense alleles. For the measurement we counted the average ratio of the number of nonsynonymous mutation to the number of synonymous mutations for each gene (pN/pS ratio). The obtained pN/pS ratio then was compared to the pN/pS ratio in genes without nonsense mutations. Genome sequences processing was performed by SAMtools, VarScan and SnpEff.

According to the obtained results, the most of synonymous mutations are located at the 3′-end of gene, where they are less harmful. Also it was shown that nonsense alleles, common for many species in population, are rare, which must be due to negative selection against them. The average pN/pS ratio appears to be about 1 for genes without nonsense alleles and slightly more than 1 (1.2) for genes with nonsense alleles. Such results means that negative selection does not act on any gene, which can not be true. The mistake could be explained by possibly wrong variant annotation.

Increasing the length of introns due to transposable elements

A. Murzina[1], I. Poverennaya[2]

[1]ITMO University, 49 Kronverksky Pr., St. Petersburg 197101, Russia
e-mail: murzinanastasiia@gmail.com
[2]Vavilov Institute of General Genetics Russian Academy of Sciences, 3 Gubkina str., Moscow 119333, Russia
email: I.poverennaya@gmail.com

Unlike exons - coding regions of a gene, intron sequences are known for a high degree of mutagenesis and, in accordance with this, great variability, so that even the length of an intron can differ greatly in related organisms. A significant increase in the length of introns may be due to the active accumulation in introns of a large number of transposable elements (TE) and repeats. In this project, our goal was to get depence between intron length and count of transposable elements.

The Dfam database is a collection of Repetitive DNA element sequence alignments. This database was used with RepeatMasker programm, which based on usage Hidden Markov Models, to search TE of human genome.

TE and repeats take up 43 percent on the average of the length of all introns. There is a correlation between the intron length and the TE length for introns with a length of more than 300 nucleotides, however, very long introns (> 12000) correlate with the TE length better than the average length introns.

Analysis of the Drosophila melanogaster full genome sequences

Anna Namyatova[1], Gennady Zakharov[2]

[1]*Saint Petersburg State University*
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russian Federation
email: anna.namyatova@gmail.com
[2]*EPAM, 22/2 A Zastavsakya ul, Saint Petersburg, 196084, Russian Federation*
email: Gennadii_Zakharov@epam.com

Drosophila melanogaster is a model object for studying insect genomes. The results can be used to make prediction on the human diseases. We had the Illumina full genome sequences for two wild type lines and two mutant lines (ts3 and X1). In the ts3 line, the defects were artificially induced with the behaviour being restored to normal after the thermal shock. The defects in the line X1 were spontaneous and permanent. The aim was to compare the mutant line genomes with each other and with those of the wild type lines, and find the genes responsible for the abnormalities in the nervous system structure and function. We used the following tools for our analysis:

1. FastQC. Sequences quality check.
2. Trimmomatic Trimming the bad quality nucleotides.
3. Bwa. Genome assembly.
4. Samtools. Creating, sorting and indexing the .bam file.
5. Picard. Adding the Readgroups into .bam files.
6. Gatk. Variation calling.
7. Vcf-merge. Merging the wild type lines mutations.
8. Rtg vcfeval. Comparing the each mutant line mutations with those in the wild type

lines.

The genomes were mapped against the reference genome

Drosophila_melanogaster.BDGP6.dna_sm.toplevel.fa.

FastQS showed that there were around 30 million reads in each genome, the length of reads ranges between 35 and 76 bp in the raw sequences. The total number of mutations in the wild type line was 1209056. Each mutant line had around 800000 mutations. There were 195743 unique mutations in the ts3 line, and 174653 unique mutations in the X1 line. In the future we are going to perform snpeff and snpsift tools to annotate the mutations, to assign the biological meaning to them and to exclude the nonsense mutations.

Automated pathway annotation for single-cell RNA-seq

M. Firuleva[1], K. Zaitsev[2]

[1]Saint Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russian Federation
e-mail: mmfiruleva@gmail.com
[2]ITMO University
Kronverkskiy pr., 49, lit. A, Saint Petersburg, 197101, Russian Federation
e-mail: zayats1812@gmail.com

Method of single-cell RNA-seq expands the opportunities to research a biological difference between cells of interest by the individual transcriptome analysis of each cell simultaneously and to study the cell's processes more deeply. Increasing pace of RNA-seq methods expects automated approaches to process a huge amount of that data. Different cell processes are mediated by a different set of genes (signal pathways), and expression of appropriate genes changes due to activation or deactivation of appropriate signal pathways.

The main target of this project is to develop a method for automated annotation of pathways which are significantly upregulated in the single-cell dataset.

We developed a three-step approach to identify differentially expressed pathways, which is applied after performing the usual single-cell rna-seq pipeline using Seurat package.

First, we calculate how each pathway is expressed in every cell. Second, randomly sampling gene sets we identify candidate cells that in which pathways are upregulated more than at random. Third, we identify clusters in which there are more candidate cells than at random, using hypergeometric distribution. As a result, our program returns a matrix with cell clusters as columns and pathways as rows which values are adjusted p-values.

Developed approach combined with cumulative statistic approaches allows to quickly find significantly upregulated pathways in a single-cell dataset for all clusters and large gene set databases.

Automated marker descriptor for single-cell RNA-seq

D. Gorbach[1], K. Zaitsev[2]

[1]Saint Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russian Federation
e-mail: daria.gorba4@yandex.ru
[2]ITMO University
Kronverkskiy pr., 49, lit. A, Saint Petersburg, 197101, Russian Federation
e-mail: zayats1812@gmail.com

Method of single-cell RNA-seq provides an opportunity to detect gene expression and specific cellular processes from lots of cells simultaneously, and each cell type has its own combination of expressed markers, which helps to discriminate one cell population from another. Increasing rate of RNA-seq technologies demands automated methods to obtain increasing amount of that data. Popular approach of cell types identification is based on “one versus all” method, which compares gene expression profiles for one cellular cluster with all the rest. It is a way to find statistically significant markers for each cluster, however, this method fails to identify “unique” cluster markers and quite often reports markers that are not unique for a certain cluster.

Cell surface markers are of particular interest for that kind of research, as they are most frequently serve as markers of specific cell types.

Our approach was to make an automated descriptor, using pair-wise comparison of expressed markers. We used MGI database for mice cell surface proteins and “Seurat” package – R toolkit for single-cell data analysis. Thus, we compared expression levels of each cell surface marker between different cellular subtypes (T_h -lymphocytes, macrophages, etc.) and obtained one or several unique markers expressed uniquely in each subtype. This method allows us to describe each cluster with a set of unique surface marker genes identifying any of that type.

SPRING 2019

Antimicrobial peptide prediction in non-model species based on transcriptome data

I. Babkina[1], L. Danilov[2], P. Drozdova [3]

[1]Saint Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
[2]Department of Genetics and Biotechnology, Faculty of Biology, Saint Petersburg State University 7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
[3]Institute of Biology at Irkutsk State University, Lenin str. 3, Irkutsk, 664003, Russia
e-mail: kriska.irichka@gmail.com

Lake Baikal hosts a unique deep-water freshwater fauna, which includes various representatives living in a wide range of environmental conditions, from the littoral zone to the maximum depth. The main diet of deep-water representatives is carrion with a specific saprotrophic microbiota. Thus, the deep-water crustaceans are assumed to have a variety of protection mechanisms against pathogens, including antimicrobial peptides (AMPs).

RNA-seq data can be used to predict the proteome, and thus they can also be used to search for AMPs. Kim et al., 2016 developed a pipeline to search for AMPs in cockroach transcripts. However, the specifics of our data and additional interest also to the cryptic AMPS require its improvement.

To search for known AMP, we performed a blastp search on the db_AMP database. In transcripts of 2 species, *Ommatogammarus flavus* and *Eulimnogammarus verrucosus*, 4 groups of proteins similar in structure to decapod crustins were found. All of them have homologous sequences within the published amphipod transcriptomes. Earlier, crustins of amphipods have never been described in detail.

For the prediction of AMPs, we used the pipeline from Kim et al., 2016 with an additional step to search for cryptic AMP. In the *Ommatogammarus flavus* transcriptome, we discovered 916 potential AMPs that need to be checked *in vitro* and *in vivo*.

Denoising of ULI-NChIP-seq data with neural networks

D. Balashova[1], O. Shpynov[2]

[1]*Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia*

[2]*JetBrains Research, Primorsky avenue, 68-70, Saint Petersburg, 197374, Russia*

Chromatin immunoprecipitation followed by sequencing of the next generation (ChIP-Seq) is a powerful method for identifying the entire genome's DNA binding sites for transcription factors and other proteins. The limitations of ChIP-seq include a large number of cells needed to create high-quality data sets. The ultra-low-input micrococcal nuclease-based native ChIP (ULI-NChIP) protocol, that was presented in the paper "An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations" (Brind'Amour J et al., 2015), requires significantly less material and usually provides a reliable peak calling, but is much more variable than the traditional ChIP-seq approach.

In the paper "Denoising genome-wide histone ChIP-seq with convolutional neural networks" (Pang Wei Koh et al., 2017) authors introduce a convolutional denoising algorithm, Coda, that uses convolutional neural networks to learn a mapping from suboptimal to high-quality histone ChIP-seq data. We analyzed the ULI-NchIP-seq data quality of histone modifications H3K27ac, H3K27me3, H3K36me3, H3K4me1 and H3K4me3 and focused on the signal-to-noise ratio (SNR) metric. We present DCNN algorithm – denoising convolutional neural network – the purpose of which is to improve the quality of the data with respect to the SNR. The essence of the method lies in the matching of high and low quality data of some histone modification, as well as, optionally, using data of other histone modifications to improve accuracy. This approach allows to transfer information from low-input noisy processes in a flexible model that can be used for noise reduction of new ULI-NChIP-seq data.

Noisy peak calling

D. Chaplygina[1], O. Shpynov[2]

[1]*Saint Petersburg State University*
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
[2]*JetBrains Research, Primorsky avenue, 68-70, Saint Petersburg, 197374, Russia*

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is a method used to analyze protein interactions with DNA. The goal of peak calling algorithm is to identify enriched areas (peaks) in a genome.

In the paper “Impact of sequencing depth in ChIP-seq experiments” (Jung et al., 2014) authors evaluated the impact of sequencing depth on peaks identification. However, signal- to-noise characteristics and its influence on peak calling algorithms were not covered. In this work we tried to estimate the impact of noise level in ChIP-seq data on enriched regions identification for core histone marks (H3K27ac, H3K27me3, H3K36me3, H3K4me1 and H3K4me3). We focused on MACS2 (Zhang, et al., 2008), SICER (Xu et al., 2009) and SPAN (novel semi-supervised peak calling algorithm by JetBrains Research) algorithms. The noise was introduced by mixing ChIP-seq and control reads with different proportions of control ranging from 0% to 90%. SICER and SPAN were used with default parameters and MACS2 with parameter --broad. False Discovery Rate (FDR) was set to 0.05 and 1E-6 to evaluate its influence on peak calling capabilities with noisy data.

The analysis of peaks dynamics demonstrated that both number of peaks and average length are decreasing with the increase of noise level. Then we compared algorithms by its stability and identified sets of peaks. We found that SPAN with FDR 0.05 is the most stable of three algorithms and higher noise level leads to lower peaks sets similarity. Investigation of FDR influence showed that more strong FDR values result in decreasing in both peak callers stability and peaks sets similarity.

Identification of pathway genes triggered differential expression profile changes.

D. Gorbach[1], E. Bakin[2], O. Stanevich[2]

[1] Saint Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
e-mail: daria.gorba4@yandex.ru
[2]Pavlov First Saint Petersburg State Medical University
L'va Tolstogo str. 6-8, Saint Petersburg, Russia
e-mail: eugene.bakin@gmail.com
e-mail: oksana.stanevich@gmail.com

Analysis of complex cellular pathways can be an issue. However, despite dozens of possible interconnections between genes in the pathway, it only requires to know the few key genes, that invokes changes in differential expression profile. In recent work we managed to find key genes that orchestrate early changes in differential expression during the Kaposi's sarcoma-associated herpesvirus (KSHV) invasion.

We used the real clinical data from cells infected with KSHV - a table of differentially expressed genes, that was previously visualized and selected in the Phantasus software. Then, we intersected these genes with pathways from the KEGG ((Kyoto Encyclopedia of Genes and Genomes) database, to choose pathways, that contain genes of interest. Selected pathways were processed using the KEGGgraph R package. We applied so called "breadth-first search", as we search for every "descendant gene" of every single gene in the pathway and compared the number of differentially expressed ones among them. Gene, that has the biggest amount of those genes (and the lesser number of non-differentially expressed descendants) is considered to be "the key gene", that initiate following changes in that part of the pathway.

All genes from KSHV-infected cells data were intersected with KEGG database (KEGG.db) and, as a result, 143 pathways possibly related with KSHV were obtained, and 6 perspective "key genes" were identified. We also proved its connection with KSHV pathogenesis (from literature data).

Improving peak calling in SPAN

E.N.Kartysheva[1], A.V.Dievskii[2]

[1]*Saint Petersburg State University*
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
[2] *JetBrains GmbH, Elsenheimerstr. 47, München, 80687, Germany*

ChIP-seq (chromatin immunoprecipitation sequencing) is one of the main methods to analyse DNA-protein interactions. It can be really helpful but it produces a lot of noisy data so the output has to be carefully preprocessed before being used. SPAN (Semi-supervised Peak ANalyzer) is a multipurpose peak caller capable processing both conventional and ULI-Chip-seq tracks.

The main goal of this study was to improve the peak caller model by adding new covariates to HMM (hidden markov model) using GLM (generalised linear model). We used ZINBA (Zero-Inflated Negative Binomial Algorithm) as a reference.

In this semester several classes were implemented for bioinf-commons such as weighted regression, emission regression scheme, poisson regression scheme and zero-poisson mixture to extend the library with methods necessary for future integration of new covariates. One of the aforementioned classes (namely weighted regression) was proposed as a pull request to Apache Commons Statistics developer branch.

In future we plan to add zero-poisson mixture to SPAN model, test it on real data and replace poisson regression with negative binomial if necessary.

Detection of pathogenic INDELs and SNPs in whole-exome sequencing data of patients with different types of idiopathic cardiomyopathy

D. Kilina[1,2], A.Kiselev[2]

[1] Preclinical Translational Research Centre
43 Dolgoozernaya Street, Saint Petersburg 197343, Russia
e-mail: dasha-kilina@mail.ru
[2]Almazov National Medical Research Centre
2 Akuratova Street, Saint Petersburg 197341, Russia
e-mail: artem.kiselyov@gmail.com

Idiopathic cardiomyopathy is a primary cardiovascular disease with high heterogeneity caused by functional lesion in cardiomyocytes in consequence of genetic abnormalities. In this study, we undertook a whole exome sequencing (WES) approach to identify novel candidate single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) associated with different types of idiopathic cardiomyopathies.

The cohort consisted of 10 patients with idiopathic cardiomyopathy. This samples were processed with Variant Calling pipeline (GATK) implemented on Snakemake and annotated by annovar, snpEff. The total number of SNPs - 203 827, insertions - 14 352, deletions – 19 891. Among discovered SNP 48.07% variants were missense, 51,38% variants were silent and 0,55% were nonsens. As a result of comparing the allele frequencies of observed variants against all ExAC population and taking into account the tool prediction (SIFT, Polyphen2, LRT, MutationTaster, MutationAssessor, FATHMM, Provean, CADD, MetaSVM) 6 previously unknown SNPs with damage effect were detected. New INDELs variants haven’t been identified.

Mutation in SLC2A8 gene can lead to increased reliance on glucose utilization contribute to the development of cardiac dysfunction, but we cannot state it with certainty on this stage of the project. Hence, the effect of discovered mutations on the pathogenesis of idiopathic cardiomyopathy needs further investigation.

Diversity of opsins in transcriptomes of Baikal endemic amphipodes

A. Kizenko[1], Y. Fedorova[2], P. Drozdova[3]

[1]*Saint-Petersburg State University*
7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia
[2]*Skolkovo Institute of Science and Technology*
3 Nobel Street, Skolkovo, Moscow, 143026, Russia
[3]*Institute of Biology at Irkutsk State University*
Lenin str. 3, Irkutsk, 664025, Russia

Amphipoda are malacostracan crustaceans generally characterized with laterally compressed bodies. Amphipod species inhabit different areas and depths of seas and freshwater bodies. Orientation of amphipods in water is regulated by opsins, G protein-coupled transmembrane receptors, which form visual pigments together with retinal chromophores and play key roles in animal photoreception. Baikal endemic amphipods vary in habitat and color, so the aim of this project was to discover which opsin genes these amphipods have and how natural selection has influenced the representation of opsin types.

We analyzed quality of transcriptomes using BUSCO v3 (Benchmarking Universal Single-Copy Orthologs). As the quality of assemblies was rather bad, we filtered out species transcriptomes of which possessed more than 20% missing BUSCOs. Then we applied the PIA (phylogenetically-informed annotation) pipeline to the remaining transcriptome assemblies. We slightly modified this pipeline to make it more useful for genes' search in transcriptomes of bad quality and added the Gblocks step. Gblocks eliminates poorly aligned positions, which can occur due to the partial CDS alignment, so that alignment becomes more suitable for phylogenetic analysis.

Finally, we discovered that Baikal amphipods possessed only long-wave sensitive opsin genes and opsin-like proteins. We suppose that Baikal amphipods have lost the expression of short-wave and ultraviolet sensitive opsins due to natural selection. The project is available in our github repository https://github.com/AlenaKizenko/diversity_of_opsins_in_amphipods.

Whole-genome Drosophila sequence analysis - 2

A. Kosolapova[1,2], I. Lebeda[3], G. Zakharov[4]

[1] *All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, 196608, Russia*
[2]*Saint-Petersburg State University*
7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia
[3]*Pavlov First Saint Petersburg State Medical University*
6-8 L`va Tolstogo str., Saint Petersburg 197022, Russia
e-mail: 1.u.r.o.n.3@gmail.com
[4]*EPAM systems, Zastavsakya ul, 22/2 A, 196084, Russia*
email: Gennadii_Zakharov@epam.com

Drosophila fruit fly (D.melanogaster) is a widely known and popular model organism. Due to the small size of the Drosophila genome and its rapid reproduction, it is fairly easy to study various conditions and diseases. In this work, DNA sequences of two mutant D.melanogaster strains (X1 and ts3) with disturbances in the structure and functioning of the nervous system, and three wild-type strains without physiological pathologies were studied. Both of these mutant strains can potentially be used to study Williams syndrome in humans.

After the initial quality control of the Illumina NGS reads, they were aligned to the reference genome, and variations were found in each strain. After that, from the list of variants of mutant strains, variants of wild-type strains were removed. In the region of limk1 gene (potential origin of variations forming mutant phenotype according to previous studies), several variations were found that were identical for both mutant strains; however, after filtration these variants were eliminated. The final set of variations unique for mutant strains was divided into groups by significance for manual analysis. The entire pipeline was performed in Snakemake and can be accessed at https://github.com/IuriyLeb/drosophila_project.

The influence of molecular dynamics parameters on protein motion characteristic timescale

O.O. Lebedenko[1], S.V. Legkovoy[2], S.A. Izmailov[3], N.R. Skrynnikov[3,4]

[1]Peter the Great Saint Petersburg Polytechnic University 29 Polytechnicheskaya str., Saint-Petersburg, 195251, Russia
[2]Saint-Petersburg State University 7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia
[3]Laboratory of Biomolecular NMR, St. Petersburg State University, St. Petersburg, 199034, Russia
[4]Department of Chemistry, Purdue University, 560 Oval Drive, West Lafayette, IN, 47907-2084, USA

Molecular dynamics (MD) modeling of biomolecules is one of the most important and promising tools of structural biology. As a rule, in MD simulations of biomolecules use the so-called NPT ensemble. However, the use of NPT with a standard set of parameters for barostat and thermostat may lead to significant changes in the characteristic time scale of the simulated motions.

The main objective of this project was to find optimal parameters for correct representation of motional timescales in two structurally dissimilar proteins: globular protein ubiquitin (ubq) and intrinsically disordered N-terminal fragment of histone H4 (h4). In this work, we recorded and processed a number of MD trajectories using different water models (spce, tip3p, tip4p-d, tip4p-ew) and ensembles (NVE, NVT, NPT γ=0.01, NPT γ=2) under Amber ff14SB force field. Based on these trajectories, we calculated the characteristic times of rotational motion for globular ubq protein. Furthermore, the characteristic times of translational diffusion for ubq and h4 have also been calculated.

As a next step, we compared all MD-derived correlation times with the corresponding experimental values. It was found that certain combinations of water models and statistical ensembles correctly reproduce the rotational diffusion process (overall tumbling): spce / NVE, tip4p-ew / NVE, tip4p-ew / NPT (γ = 0.01), tip4p-d / NVE. However, translational diffusion requires further investigation since none of the attempted procedures produced the correct results for both ubq and h4.

We would like to thank Prof. D.A. Case for drawing our attention to this problem.

Association rule mining using fishbone diagrams

N.B.Lukashina[1], D.V.Likholetova[2], P.S.Tsurinov[3], O.Y.Shpynov[3]

[1] Peter the Great Saint Petersburg Polytechnic University
29, Polytechnicheskaya str., Saint-Petersburg, 195251, Russia
[2]Saint-Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia
[3]JetBrains Research, Primorsky avenue, 68-70, Saint-Petersburg, 197374, Russia

Exploiting associations is central in human reasoning and decision making. There is a lot of rule extraction techniques from observational data. Association Rule Mining (ARM) is one of the most popular methods. Fishbone ARM (FARM) is a new data mining algorithm for constructing hierarchical associations, which can be visualized as Ishikawa diagrams. FARM was developed by JetBrains Research, and our tasks were to create a web service and validate the method on real biological datasets.

In this study we improved method usability and created the web application which allows to go from raw data to visualization in web browser. Several other data-mining algorithms were implemented: FP-growth ARM and Decision Tree. Comparison of FARM with existing methods showed its superior clarity in reporting results.

We applied FARM to data from the study “A Validated Regulatory Network for Th17 Cell Specification” (Ciofani et al., 2012). Authors used genome-wide TF occupancy, expression profiling of TF mutants, and expression time series to delineate the Th17 global transcriptional regulatory network, identifying multiple new Th17 regulators. FARM was able to reconstruct main complexes of TFs acting in the Th0-Th17 differentiation process.

After FARM validation on genome data we focused on biochemistry assays of blood and urine datasets to construct fishbone diagrams for old and young patients, and found associations of creatinine clearance and TNF-alpha receptor I, previously reported to be connected with age (Ogna et al., 2015, Schaap et al., 2009).

Analysis of yeast genomes from the Peterhof genetic collection

A. B. Matiiv[1], Y.A. Barbitoff[1,2], A.V. Predeus[2,3]

[1]Saint-Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia
[2]Bioinformatics institute
2A Kantemirovskaya, Saint Petersburg, 197342, Russia
[3]University of Liverpool, UK, Liverpool, L69 3BX

The Peterhof genetic collection of Saccharomyces cerevisiae strains (PGC) is a large laboratory fund, which has accumulated several thousand strains for more than half a century. Several PGC strains have been widely used in certain areas of yeast research, but their genomes have not yet been fully studied. The genetic distance between the precursor PGC and S288C is comparable to that between two geographically isolated populations. This project is a continuation of a project to assemble the yeast genome from Oxford Nanopore (ONT) data. During this project it was supposed to identify compensatory mutations that allow cells to survive with the disruption of the gene encoding the vital translation termination factor.

At the moment we have obtained 3 genome assemblies of *Saccharomyces cerevisiae* 1A-D1628 strain: (1) draft genome assembly from ONT reads, that was assembled previously with canu, (2) enhanced polished with Nanopolish genome assembly, and (3) even more enhanced with Racon and Nanopolish genome assembly. If compared with QUAST, the final (3) assemble had 205.12 mismatches per 100 kbp (against 212.37 for (1) and 207.85 for (2)) and 25.04 ndels per 100 kbp (against 343.76 for (1) and 73.09 for (2)). So, we can assume enhanced with Racon and Nanopolish genome assembly is our best genome assembly of *Saccharomyces cerevisiae* 1A-D1628 strain.

In order to make annotation of *Saccharomyces cerevisiae* 1A-D1628 strain we used Exonerate, Maker, RGAAT and RAAT tools. But only with Exonerate it was possible to create annotation with similar gene numbers compared with closely related *Saccharomyces cerevisiae* S288C strain. Also, this obtained annotation was used for SnpEff database building.

Towards detection of differential RNA editing events in transcriptomics datasets

A. Matveenko, A. Samsonova & A. Kanapin

Saint-Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia

Single-nucleotide modifications of RNA, or RNA editing, is an important regulatory mechanism in the cell. However, understanding of its regulation is far from complete, as transcriptome-wide quantitation of the RNA editing is complicated and requires developement of new computational approaches. Most tools for RNA editing analysis are limited to the search of potential editing sites and do not support the analysis of differential editing (DEd). On the other hand, great number of approaches exists for analysis of differential methylation (DM), specifically, bisulfite sequencing, which is similar in data modality to the RNA editing. In this project we aimed to evaluate applicability of existing tools and statistical approaches for analyses of bisulfite sequencing data, for discovery of DEd events.

We attempted to identify DEd sites and genes in the RNASeq data of BT20 cell line. The samples were prepared either under hypoxia or in normoxia, three replicates per condition. Previously, in a work by Irina Shchukina, a tool for discovery of A to I RNA editing events in RNASeq data was developed. The output of this tool listing editing sites determined in the cell line was used as the data for the subsequent analysis. edgeR pipeline for analysis of DM was modified here to apply for the Ded analysis both at single-nucleotide and at gene level. As the result we obtained lists of DEd sites and genes in the sample, and found that RNA editing of approximately 100 genes is enhanced under hypoxia. GO-enrichement analysis of the gene list revealed that RNA editing is enhanced during hypoxia in genes acting in ribosome biogenesis, mitochondrial translation, and transcriptional regulation associated with hypoxia. Thus, edgeR DM pipeline can be used for differential RNA editing analysis.

Searching for latent viruses in human whole genome sequencing data

A.Morshneva[1], N.Pogodina[1], I.Orlov[2], A.Rakitko[3], V.Ilinsky[3]
[1]*Saint Petersburg State University*
7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia
e-mail: 1195alisa@gmail.com, nadezda.pogodina.bio@gmail.com
[2]*ITMO University 49 Kronverksky Pr., St. Petersburg 197101, Russia*
e-mail: orlov239@gmail.com
[3]*Genotek 7/1 14 Nastavnicheskiyi per., Moscow 105120, Russia*
e-mail: valery@genotek.ru

Viral ability to stay in an asymptomatic phase of infection when a virus is not replicating is known as latency. A virus can stay latent for decades without exposing itself, although maintaining a capability to cause acute infections. Hence, viral presence in the human organism can remain undetected.

Human whole-genome sequencing data contain sequences of dsDNA-viruses (or integrated RNA-viruses) apart from human sequences, because all such sequences are technically indistinguishable in the host DNA. The viral sequences aren't aligned to the reference human genome and can be identified by mapping to the viral genomes. There are six viral families which we initially expected to observe as they store their genetic material in dsDNA form: Adenoviridae, Herpesviridae, Poxviridae and Polyomaviridae (linear dsDNA), Papovaviridae and Hepadnaviridae (circular dsDNA).

In this project we estimated the viral load in WGS data (blood samples) from private Genotek database and 1000 Genomes project. For this purpose we developed a pipeline for detecting viruses in human WGS data using Kraken2. Almost in all the samples *Epstein-Barr virus* (EBV) was found. *Mastadenoviruses* were detected only in 23% of the samples.

In order to determine an associations between viral load level of EBV/viral presence of *Mastadenoviruses* and SNPs in human genome we performed GWAS using genuine Genotek pipeline on 5 european populations from 1000 Genomes project (CEU, FIN, GBR, IBS, TSI). GWAS revealed several significant SNPs possibly connected with the viral load rate. However, their functions should be explored.

Adaptation of fish to the depth

Anna Namyatova[1], Elena Polyakova[2], Nadezhda Potapova[3]
[1]All-Russian Institute for Plant Protection, 3, Podbelskogo sh., Pushkin, St. Petersburg, 196608, Russia
[2]EPAM Systems, 11/7c4, Levashovskiy pr, St. Petersburg, 197110, Russia
[3]Faculty of bioengineering and bioinformatics, Lomonosov Moscow State University, Leninskiye Gory, Moscow, 119991, Russia

Previously it was shown that most species of deep water fishes possess a rod-only retina with a pigment that is usually shortwave (Hunt et al. 2001). Therefore, it can be suggested that there might be similar changes in the rhodopsin protein in different species, leading to the same adaptation.

To test this hypothesis we downloaded the rhodopsin sequences from Genbank belonging to 34 deep water fish species from seven orders with the recoded depth of living exceeding 3000 m. We also downloaded the rhodopsin sequences for 28 shallow water species. The aminoacid sequences were aligned in Geneious software. Based on this alignment, we calculated the frequencies of the aminoacid changes using the Python script. For comparison between the deep water and shallow water fishes we calculated the differences between the squared frequences of the each aminoacid change in deep water fish and shallow water fish. We chose the changes based on the following criteria: (1) the difference of the squared frequences should be >0.3; (2) there should be more changes in the deep water fishes than in the shallow water fishes; (3) the identical change should occur in >2 orders. Overall, six changes fitted our criteria. We also showed that all those changes were in the domain and that the replacing aminoacids were very similar to the replaced aminoacids in two positions (<30 based on the Grantham's score).

The further analysis requires checking the active sites of the protein to draw conclusion on the importance of the found changes for the protein function. Running the analysis for the larger dataset will provide more robust results.

Local sequence alignment using intra-processor parallelism

D. Orekhov[1], A. Tiskin[2]

[1]*Saint-Petersburg State University*
7-9 Universitetskaya Emb., Saint-Petersburg, 199034, Russia
e-mail: d.i.orekhov@gmail.com
[2]*University of Warwick, Coventry CV4 7AL, UK*
e-mail: A.Tiskin@warwick.ac.uk

Local alignment of DNA sequences is a fundamental problem of bioinformatics. Standard solutions include fast heuristic methods such as BLAST, as well as the more time-consuming exact methods. An efficient exact local alignment technique, based on a "sliding window" approach, was developed by a University of Warwick team, resulting in a number of biologically significant results. The efficiency of that implementation was achieved, in particular, by utilising low-level intra-processor parallelism. In recent years, commodity processor architecture has been developing rapidly, culminating with Intel's AVX-512, an instruction set taking intra-processor parallelism to a new level of efficiency and sophistication, while also being surprisingly well-suited for speeding up the "seaweed combing'' sequence alignment technique developed by the second author.

We developed a prototype software tool that is, to our knowledge, the first sequence alignment software taking advantage of AVX-512 parallelism. Our tool allows one to produce semi-local alignments between short DNA fragments and long DNA strings, using seaweed combing and intra-processor parallelism to achieve competitive performance. In future, we plan to extend our implementation to a very fast exact local sequence aligner with "sliding window" functionality.

Effect of smoking on human leukocyte epigenome

P. Pavlova[1], M. Firuleva[1], Y. Kornienko[2], O.Sergeyev[3]

[1]Saint Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia
2 Institute of Cytology RAS, 4 Tikhoretsky avenue, Saint Petersburg 194064, Russia
3 A.N. Belozersky Research Institute Of Physico-Chemical Biology, Moscow State University
119234 Leninskye Gory, House 1, Building 40, Moscow, Russia
e-mail: olegsergeyev1@yandex.ru

Environmental factors, including chemicals, can cause epigenetic changes that can be traced to subsequent generations. The most studied epigenetic changes are DNA methylation, small non-coding RNA, and histone modification. Smoking remains one of the most adverse voluntary health risks. Reduced representation bisulfite sequencing (RRBS) data can be used to study the pattern of methylation changes upon exposure to smoking.

Our project is a part of the Russian Children's Study, a prospective cohort of 516 boys who were enrolled at 8–9 years of age and provided semen and blood samples at 18–19 years of age (Sergeyev et al, 2017). We analysed smoking influence on the DNA methylation level of peripheral blood leukocytes at the age of 18. To search for differentially methylated CpG islands and regions (DMR), we used two different approaches.

To implement the first approach, we used data from the CpG islands presented in all samples of peripheral blood. After exclusion regions that did not meet the inclusion criteria, we used the A-clustering algorithm (Sofer et al, 2013) to combine the regions into clusters and generalized estimating equation model to search for significant DMRs. With this approach we identified 217 A-clusters, from them 77 were significant (p-value < 0.05). Because for A-clustering implementation we needed to restrict our data, a lot of important information could have been lost.

DMRcate - R package for search of differentially methylated regions (DMRs) associated with exposure to a factor (Peters at al, 2015). In our project, we used smoking past half year in binary classification (smoke or not) to find DMRs associated with exposure. 145 significant CpGs and 34 significant DMRs (p-value < 0.05) were found in our data. From them 19 DMRs overlap with at least one promoter (reference - GRCh 38). We found 23 genes associated with significant DMRs. These genes are associated with antisense RNAs, lincRNAs, miRNA, pseudogenes, zinc fingers, transcriptional factor, spliceosome, cell

adhesion and migration, kinase, metalloprotease, electron transport chain and amino-acid transporter.

Finally, we found DMRs using two different statistical strategies for analysis of DNA RRBS data. Further research plans include the analysis of changes in the expression of various groups of small RNAs, as well as a comparative analysis of leukocyte and semen RRBS data. The project is available in our github repository: https://github.com/mariafiruleva/leykocytes_dmr_analysis.

Detection of CNVs in patients with different types of idiopathic cardiomyopathies

O.V. Romanova[1,2], A.M. Kiselev[3]

[1]The St.Petersburg State Health Care Establishment the City Hospital №40, Russia, Sestroretsk, Borisova str., 9, 197706
[2]Bioinformatics institute, Russia, Saint Petersburg, Kantemirovskaya, 2a, 197342
[3]Federal Almazov Medical Research Center, Laboratory of Molecular Biology and Genetics, Russia, Saint Petersburg, Accuratova street, 2, 197341

Array comparative genomic hybridization (aCGH) is considered the gold standard for copy number variation (CNV) detection. However, next-generation sequencing (NGS) is developing rapidly. To date there are a lot of NGS laboratories which are using analysis of whole-exome sequencing in the medical diagnostic of inherited diseases. Despite of NGS technology is widely disseminated, it is generally not used for CNV detection. We believe that using NGS to identify CNVs and SNVs could be of benefit to laboratories saving time and reducing costs while creating a more comprehensive picture of genomic variation with a single assay.

An aim of this study was to detect of CNVs in whole-exome sequencing data of patients with different types of idiopathic cardiomyopathies. To achieve our goal, we realized two pipelines CNVkit (https://github.com/etal/cnvkit) and ClinCNV (https://github.com/imgag/ClinCNV) using snakemake, analyzed a cohort of 10 patients' whole-exome sequencing data using developed scripts and carried out research clinically relevant structural variants in patients with different types of idiopathic cardiomyopathies.

With regard to findings of the study there was detect a chromosomal 15q partial deletion (q11.2-q13.1) in one of the patients. This structural variation was verified by aCGH. According the literature, the loss of this region associated with Prader-Willi syndrome (PWS).

Bayesian optimization for demographic history inference

I.V. Sheshukov[1], E.E. Noskova[2], V.A. Borovitskiy[3]

[1] *Mathematics and Mechanics Faculty, St. Petersburg State University, University emb., 7-9, St. Petersburg, 199034, Russia*
[2]*Computer Technologies Laboratory, ITMO University, St. Petersburg, Russia*
[3]*Chebyshev Laboratory, St. Petersburg State University, 14th Line V.O., 29B, Saint Petersburg 199178, Russia*

Demographic histories are studied in population genetics to infer the way populations migrate, split and change its size.. One of the most used tools in this area is moments.

The project goal was to replace a BFGS (Broyden Fletcher Goldfarb Shanno) optimization algorithm used in the moments tool with the Gaussian process based global Bayesian optimization and to study the effects. Bayesian optimization is a family of global optimization algorithms which can be more appropriate for a given task.

As a result, we successfully integrated optimization routine from the library GPyOpt into moments. Then we compared the results: our non-exhaustive tests showed that our solution was converging faster than the moments library. Later more exhaustive testing needs to be done.

Implied weighting as a measure of clade support: automation of the task and comparative assessment of results.

E.S. Sytnik[1], L.G. Danilov[2], F.V. Konstantinov[3]

[1]Peter the Great Saint Petersburg Polytechnic University 29 Polytechnicheskaya str., Saint-Petersburg, 195251, Russia
[2]Department of Genetics and Biotechnology, Faculty of Biology, Saint Petersburg State University, 7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia
[3]Department of Entomology, Faculty of Biology, Saint Petersburg State University 7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia

Implied weighting (IW) is a method in parsimonial phylogenetic analysis that allows to assign weights to characters in data according to their appearance in homoplasies and additional parameter k. k - parameter that describes the degree of concavity of the function, which describes the weight of signs This method is usually used for morphological data for which some characters are more important than others. This is important for parsimonial analysis and assigning weights can significantly affect the resulting trees.

The most common program for IW analysis is TNT which was developed by Goloboff. It has a default k = 3 which is considered not recommended due to eliminating of homoplasies at lower k values. At the same time there is no consensus about optimal k value, usually only one value is used.

We have used the idea of applying IW with different k values as a clade support method as more stable clades should be obtained in the wide range of k values. We have constructed TNT and Python scripts allows to set minimal and maximum for k and calculate optimal k values based on logarithmic scale. Initial tree and IW are done using command-line TNT. The output of the scripts is a majority-consensus tree with bootstrap values calculated as a percentage of k values where this clade is presented. Resulting tree can be received is user-defined format (nexus, newick, phyloxml, nexml, cdao).

Systematic comparison of state-of-the-art variant callers' performance

R. Abasov[1], V. Tvorogova[2,3], A. Shikov[3], Y. Barbitoff[2,3]

[1] Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, 117997, 1 Samory Mashela str., Moscow, Russia

[2]Saint Petersburg State University, 199034, 7/9 Universitetskaya emb., Saint Petersburg, Russia

[3]Bioinformatics Institute, 197342, 2A Kantermirovskaya st., Saint-Petersburg, Russia

Exome sequencing is the technology of sequencing all protein-coding parts of the genome, i.e.,sequencing of all exons. Exome sequencing is widely used in medical genetics as most of the relevant variations are concentrated in the coding loci. The processing of raw exome sequencing reads generally includes read alignment, variant calling, and filtering of the identified variants (single-nucleotide polymorphisms (SNPs) and insertion-deletion variants (indels)) using various methods. The output of such pipeline represents a VCF file containing a list of variants identified in one or several samples. When analyzing the human genomes, especially when searching for variants associated with disease, the accuracy of variant calling is extremely important. Our goal was to evaluate this accuracy for different variant calling pipelines.

The main goal of this project was to compare different pipelines for exome variant calling. We used the following tools in our analysis: GATK3, GATK4, STRELKA v. 2.9.10, DEEPVARIANT v. 0.7.2, and FREEBAYES v. 1.2.0. For evaluation of the accuracy of small variant discovery we used the gold standard datasets from the Genome In A Bottle (GIAB) Project, as well as the recently developed Synthetic Diploid Sequence (SynDip) benchmark dataset. We compared the average quality of the variant calls across all samples using the F1 metric that is widely used to evaluate the performance of classification algorithms. For small indels, the older version of GATK - GATK3 showed the best variant call quality, unexpectedly outperforming the newer version of the tool, GATK4. For SNPs, DEEPVARIANT showed the best results. We also found that the variant call quality is highly variable among different samples even for the same variant calling tool, supposing that factors other than variant caller, such as sequencing coverage, sample preparation, and type of experiment (exome or genome sequencing), also influence the accuracy of small variant discovery.

SPAdes support for third-party assembly graphs

N.Zenkova[1], A.Korobeynikov[2], A.Prjibelski[2]

[1]Saint Petersburg State University, 7/9 Universitetskaya Emb., St. Petersburg 199034, Russia
[2]Center for Algorithmic Biotechnology, St. Petersburg, 199004, 6 linia V.O.. 11/21d, Russia

SPAdes – St. Petersburg genome assembler – is an assembly toolkit containing various assembly pipelines. SPAdes outputs contigs and genome assembly graph in FASTG and GFA formats, but do not have the ability to accept such graphs as input. FASTG is a format for faithfully for representation of genome assemblies in the face of allelic polymorphism and assembly assembling uncertainty. The Graphical Fragment Assembly (GFA) is a tab-delimited text format for describing a set of sequences and their overlap. The goal of this research project was implementation of support for third-party assembly graphs, in particular, graph gfa format. The implementation of this functionality significantly expands using SPAdes. It allows using repeat resolution and scaffolding based on graphs and built by other genome assemblies. For this implementation the new stage load_graph replacing stages construction and simplification was constructed. For downloading graph class LoadGraph was created. At the moment, SPAdes accepts GFA graph from load_graph stage with Load function. Coverage for edges are read from GFA file.

There is another problem: SPAdes accepts de Bruijn graph as input and do not have the ability to input another type of sequence graph with overlap equal to zero. The simpler algorithm was proposed for solving this problem. The goal is transforming arbitrary graph to deBruijn graph. This algorithm consists of three stages: iteration through vertices, continuation outgoing edges and complementary edges and repeat until all vertices are processed. But this method is not universal. It is necessary to modificate this algorithm for some cases: e.g. for situations where there are equal edges, edges shorter than k, where k is k-mer size, in input graph or there is an vertex containing different incoming and different outgoing edges.

SUMMER 2019

Identification of pathway genes triggered differential expression profile changes.

D. Gorbach[1], E. Bakin[2], O. Stanevich[2]

[1]Saint Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia
e-mail: daria.gorba4@yandex.ru
[2]Pavlov First Saint Petersburg State Medical University
L'va Tolstogo str. 6-8, Saint Petersburg, Russia
e-mail: eugene.bakin@gmail.com
e-mail: oksana.stanevich@gmail.com

Analysis of complex cellular pathways can be an issue. However, despite dozens of possible interconnections between genes in the pathway, it only requires to know the few key genes, that invokes changes in differential expression profile. In recent work we managed to find key genes that orchestrate early changes in differential expression during the Kaposi's sarcoma-associated herpesvirus (KSHV) invasion.

Previously, we used the real clinical data from cells infected with KSHV - a table of differentially expressed genes, that was visualized and selected in the Phantasus software. This time, we were provided with data on influenza infection on PBMC (peripheral blood mononuclear cells) and were able to try our approach and to use some of new ideas we came up with. Also, we used already published data – study on presymptomatic detection of infection in humans (Influenza H1N1 and H3N2) (Woods et al., 2013) . All samples were divided by the virus type and clustered by the time of virus inoculation and onset of symptoms (H1N1, 93.5 hours after inoculation, H3N293.5 hours after inoculation, symptomatic sub-group; we also tried to compare symptomatic and asymptomatic sub-groups with same parameters). Thus, not only are we discovered specific pathways and "key genes" for the infection development, but also identified differences in that process for two types of influenza virus depending on the duration of the illness.

First, genes from groups were intersected with pathways from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database, to choose pathways, that contain genes of interest. Selected pathways were processed with the KEGGgraph package, submitting every pathway in a form of oriented graph.

Then, we applied so called “breadth-first search”, as we search for every “descendant gene” of every single gene in the pathway and compared the number of differentially expressed ones among them. After processing all graphs separately, we merged them into one “mega-graph” and applied the same approach on it.

Gene, that has the biggest amount of differentially expressed descendants (and the lesser number of non-differentially expressed ones) is considered to be “the key gene”, that initiate following changes in that part of the pathway. Finally, we minded intersections between pathways in the mega-graph and genes, presented in two or more pathways, considered to be the “key genes” of particular interest for that study.

Thus, we discovered sets of genes for every group, mentioned above, and examined received information. Publication results partly corresponded to ours (5 out of 20 from each set), which may be the validation of our approach. In particular, we found certain correlations between onset of symptoms and up-regulation of proteasomal genes, moreover, immunoproteasome genes. In addition, significant amount of genes were overexpressed in the RIG-I-like receptor signaling pathway (Loo et al., 2011) , and one gene was defined as “key” - ISG15 (interferon-stimulated gene 15). Further investigation showed its crucial role in influenza pathogenesis and also its potential role in therapy (Zhao et al., 2016) .

1. Loo, Y.-M., & Gale, M. J. (2011). Immune signaling by RIG-I-like receptors. *Immunity*, 34(5), 680–692.

2. Woods, C. W., McClain, M. T., Chen, M., Zaas, A. K., Nicholson, B. P., Varkey, J., … Ginsburg, G. S. (2013). A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2. *PloS One*, 8(1), e52198.

3. Zhao, C., Sridharan, H., Chen, R., Baker, D. P., Wang, S., & Krug, R. M. (2016). Influenza B virus non-structural protein 1 counteracts ISG15 antiviral activity by sequestering ISGylated viral proteins. *Nature Communications*, 7, 12754.

Improving peak calling in SPAN

E.N.Kartysheva[1], A.V.Dievskii[2]

[1]*Saint Petersburg State University*
7/9 Universitetskaya Emb., Saint Petersburg, 199034, Russia
[2]*JetBrains GmbH, Elsenheimerstr. 47, München, 80687, Germany*

ChIP-seq (chromatin immunoprecipitation sequencing) is one of the main methods to analyse DNA-protein interactions. It can be really helpful but it produces a lot of noisy data so the output has to be carefully preprocessed before being used. SPAN (Semi-supervised Peak ANalyzer) is a multipurpose peak caller capable processing both conventional and ULI-Chip-seq tracks.

The main goal of this study was to improve the peak caller model by adding new covariates to HMM (hidden markov model) using GLM (generalised linear model). We used ZINBA (Zero-Inflated Negative Binomial Algorithm) as a reference.

In this summer I created beta-version of SPAN with mixture of two Poisson regressions and zero-inflated component, optimize it and test on real data. Because of absence of Markov model our model make more peaks. It’s not always good so as next step we want to create Markov model with poisson regression and maybe change poisson regression to negative-binomial because negative-binomial distributions better fits real data.

Searching of potential markers of pregnancy complications using sequencing data of circulating cfDNA from maternal plasma

A. Morshneva[1,2], P. Kozyulina[3], A. Glotov[3]

[1]Saint Petersburg State University
7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia
e-mail: 1195alisa@gmail.com
2 Bioinformatics Institute
2A Kantemirovskaya, Saint Petersburg 197342, Russia
3 D.O.Ott's Research Institution of Obstetrics, Gynecology and Reproductology,
3 Mendeleyevskaya line, Saint Petersburg 199034, Russia
e-mail: polykoz@gmail.com, anglotov@mail.ru

Small DNA fragments circulating in blood and other body fluids form a pool of extracellular DNA, also known as cell-free DNA (cfDNA). During pregnancy, fetal DNA derived mostly from the apoptotic cytotrophoblast enters the bloodstream and mixes with maternal cfDNA.

Analysis of cfDNA is already widely used in prenatal screening for identifying possible chromosomal aberrations in fetus. Fetal DNA is analysed by extraction of short cfDNA fragments from low-molecular fraction, with maternal cfDNA as a by-product. Examination of both maternal and fetal cfDNA in samples can make these tests more informative since cfDNA has the potential ability to be used for the diagnosis of maternal diseases and conditions. Thus, the study aimed at exploring the possible applications of cfDNA data for clinical purposes.

The three main lines of research were identified based on literature data. First of all, the viral sequences in samples can be analysed. The developed Kraken2 based pipeline has revealed the presence of human viral reads in a number of samples, but in general, they were represented in a low number of reads (1-10). Viral reads are mostly presented in a high molecular fraction of DNA, so cfDNA can be only used for preliminary diagnostic. Nevertheless, all samples in some launches of sequencing showed large number (up to 600) of reads belonging to the particular bacterial viruses, which is an indication of the contamination of labour kits and can be used to check their quality.

Also, we tried to call SNP from cfDNA of maternal plasma. Low and incomplete coverage, which is a common problem for cfDNA sequencing, has become the main limitation here. None of the chromosomal SNPs detected had

enough depth for consideration, but the high coverage of mitochondrial DNA allowed us to detect mitochondrial SNPs, which can be used for diagnostics of clinical syndromes involving mtDNA.

The ratio between nuclear and mitochondrial DNA fragments has been reported to differ considerably between health and pathology, so the examination of this ratio has become the third line of our research. We showed that this ratio is generally lower in some pathologies, but methodological limitations do not yet permit a reliable interpretation of these results.

The results obtained will be used in further to extend the functionality of non-invasive prenatal tests.

Bayesian optimization for demographic history inference

I.V. Sheshukov[1], E.E. Noskova[2], V.A. Borovitskiy[3]

[1] *Mathematics and Mechanics Faculty, St. Petersburg State University, University emb., 7-9, St. Petersburg, 199034, Russia*
[2]*Computer Technologies Laboratory, ITMO University, St. Petersburg, Russia*
[3]*Chebyshev Laboratory, St. Petersburg State University, 14th Line V.O., 29B, Saint Petersburg 199178, Russia*

Demographic histories are studied in population genetics to infer the way populations migrate, split and change its size. A number of tools are used in this area: moments, dadi and GADMA, but they all can be improved.

This summer project continues semester project of the same name. Its goal was to extend functionality of tool we made during the previous semester, extensively test it on more datasets and if possible improve quality of optimization.

We've added new features such as: automated report generation, multidimensional plotting capabilities, time constraints and more.

We've conducted more experiments on old and new datasets with different optimization hyperparameters — the results were comparable with dadi and moments. We've also created a synthetic dataset of 5 populations and compared our tool and GADMA. Our tool converged to optimum faster than GADMA.

In the future we hope to conduct more tests and extend the capabilities of the program.

BIOINF.ME