ИНСТИТУТ БИОИНФОРМАТИКИ 2021/22

Сборник тезисов весеннего семестра

Санкт-Петербург

2022

ISBN 978-5-7422-7814-6
УДК 004.94, 575.1, 575.8, 577.2, 578.5, 579.25

ИНСТИТУТ БИОИНФОРМАТИКИ

Результаты НИР 2021/22 учебного года

Санкт-Петербург, 2022

BIOINFORMATICS INSTITUTE 2021/22

Spring term research projects

Saint Petersburg

2022

ISBN 978-5-7422-7814-6

BIOINFORMATICS INSTITUTE

2021/22. Projects Abstracts

Saint Petersburg, 2022

Table of contents

SPRING 2022

УДК 577.22

Analysis of the structural diversity of β-arches

R. Basyrov [1], L. Zhozhikov [2], S. Bondarev [3]

[1] Moscow Aviation Institute, Volokolamsk highway 4, 125993, Moscow, Russia
[2] Institute of Medicine, North-Eastern Federal University, Kulakovsky st. 42, 677007, Yakutsk, Russia
[3] Saint-Petersburg State University, Universitetskaya emb. 7/9, 199034, Saint-Petersburg, Russia

β-arches are the common structural element of numerous amyloid aggregates. These aggregates possess specific cross-β structure. Amyloids were discovered as pathological protein deposits associated with different human diseases. According to recent data, amyloid fibrils formed by a long (over about 30 residues) amyloidogenic peptides that are prone to have β-arches. β-arches include two β-strands united by a turn (β-arc) between them [1]. In amyloids and β-solenoid proteins, β-arches stack in-register to form β-arcades. Kajava's group proposed the topological classification of β-arches according to the conformation of β-arcs [2]. In the current research, we aimed to analyze the diversity of β-arches organization and to classify β-arches based on their 3D structure. Data on a total 1324 β-arches, which made up 319 types of arches, were taken for processing. The information about known β-arches was provided by A.V. Kajava. After filtration and exclusion of not representative types of arches ($n < 5$), a total 880 β-arches with 17 types of them remained. A script was written to align two β-arches and calculate RMSD (root-mean-square deviation) of atomic positions using python3. To describe the three-dimensional organization of β-arches for all possible structures, clustering of β-arches based on torsion angles of amino acids in β-strands using DBSCAN for python3 and on RMSD using hierarchical clustering with R was performed [3,4]. These parameters were calculated for the atomic positions of the first three and the last three Cα atoms of β-strands, because such residues are present in all β-arches.

The DBSCAN clustering of torsion angles gave no significant results - only two clusters were obtained, and the remaining β-arches were marked as noise. Also, a pairwise comparison of angles using the Wilcoxon signed-rank test showed their weak diversity depending on the type of arch. Further, the dendrogram was obtained as a result of hierarchical clustering by RMSD. It showed similarity with the clusters proposed in the previous articles. But also, breakdown of some clusters into smaller groups was observed. Hence, it can be concluded that RMSD, in contrast to torsion angles, is well suited for assessing the structural diversity of β-arches.

Presented results have potential implementation in the development of amyloid fibril assembly software, in sequence-based detection and structural prediction of other β-solenoid proteins, for identification of amyloidogenic sequences and

elucidation of amyloid fibril structures. Amyloidosis is a fairly common disease in the medical practice of neurological, cardiological, nephrological and endocrinological services, the diagnosis of these conditions is at a low level. Further progress made in the classification and prediction of such structures will help researchers from medical practice to elucidate approaches in the diagnostics and in the speculative future in finding the treatment of these conditions.

Detailed description of the analysis is available at the Bitbucket repository: stanislavspbgu/fibrils_3d/.

References

1. Hennetin J. et al. Standard conformations of β-arches in β-solenoid proteins //Journal of molecular biology. – 2006. – T. 358. – №. 4. – C. 1094-1105.
2. Kajava A. V., Baxa U., Steven A. C. β arcades: recurring motifs in naturally occurring and disease-related amyloid fibrils //The FASEB journal. – 2010. – T. 24. – №. 5. – C. 1311-1319.
3. DeLano WL (2002) The PyMOL molecular graphics system. http://www.pymol.org
4. Pedregosa F. et al. Scikit-learn: Machine learning in Python //the Journal of machine Learning research. – 2011. – T. 12. – C. 2825-2830.

УДК 004.94, 577.22

"Split" repeat resolution for long reads

G. Bukley[1], D. Antipov[2]

[1] National Research University Higher School of Economics, Pokrovskij bul'var 11, 109028, Moscow, Russia
[2] Center for Algorithmic Biotechnology, St. Petersburg St ate University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia

De Bruijn graphs are widely used in genome assembling problems. These graphs are built on the basis of sequencing of the genome. However, some of the information from the original reads remains unused in the graph.

Due to inaccuracies of reads and insufficient read length, unresolved repeats occur in the de Bruijn graph. Different assemblers try to resolve repeats using additional information from the reads. Widely used SPAdes [1] genome assembler uses a method based on iterative expansion of paths in a graph supported by reads.

Recently, another method for resolving repeats, Multiplex de bruijn graph [2] in the LJA assembler, has been proposed. Previously, SPAdes used a different method of resolving repeats, based on the idea of splitting vertices (split). Split was quite inconvenient to work with paired reads of Illumina. However, with long Hi-Fi reads this method seems potentially more powerful than Multiplex De Bruijn graph.

The purpose of this project was to implement the "split" repeat resolution method and compare it with Multiplex De Bruijn graph. As a result, two approaches were introduced: “Split one-to-many” and “Explicit split”. If a vertex has one incoming edge and several outgoing edges, then it can be splitted into several, — one copy for each outgoing edge (one-to-many). If the reads path through the incoming and outgoing edges from the vertex can be unambiguously splitted, then a vertex should be splitted with a copy for each such path through the vertex (explicit).

However, this is just the beginning and other methods have not yet been implemented. There are other ideas, hence there is still a hope that the “split” approach will surpass multiplex De Bruijn graph. Further development of the work may be the implementation of more advanced methods of vertex split. After that, it will be possible to compare the results obtained with the Multiplex De Bruijn graph method. If the results, as expected, turn out to be better, then this module can be rewritten from Python to C++ and be implemented in LJA assembler.

References

1. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm

and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455-77.

2. Bankevich, A., Bzikadze, A.V., Kolmogorov, M. et al. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. Nat Biotechnol (2022).

УДК 577.22

Age patterns in gene regulatory networks

Y. Burankova[1], E. Zhivkoplias[2]

[1] *Bioinformatics Institute, Kantemirovskaya street, 2A, 197342, Saint-Petersburg, Russia*
[2] *Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Solna, Sweden*

Introduction

One of the problems in systems biology remains the lack of understanding of the large-scale biological relationships between genes and the proteins they encode. The wide availability of system-level gene expression datasets makes it possible to reconstruct hidden regulatory relationships between gene-gene and gene-protein, or to reverse-engineered gene regulatory networks (GRNs) [1]. GRN comprises nodes (the genes and their regulators) and edges (the regulatory relationships between the nodes). It is usually represented mathematically as an oriented graph. The nature of the interactions in GRNs distinguishes it from other networks in biological systems. The interactions between molecules in GRNs usually involve the indirect regulatory interaction through the biological molecules, which are hard to detect and quantify. Consequently, GRNs are harder to validate.

The GRNs we know are the result of a long biological evolution. The phylogenomic analysis makes it possible to classify genes based on the oldest species that carry orthologous genes [2, 3]. For protein-protein interaction (PPI) networks in yeast and human, it was shown that proteins of the same age tend to interact more [4, 5].

This project aims to explore if gene interaction preference for genes of similar age holds in gene regulatory networks, particularly in those that describe direct regulatory interaction (transcription factor-target gene). Existing network prediction methods rely primarily on expression data. If gene interaction preference for genes of similar age holds in gene regulatory networks, incorporating biological knowledge into network inference methods could help to improve the reliability of the GRNs inferred from expression data.

Materials and methods

For the analysis, we used three gene regulatory networks. Yeast GRN is a complete transcriptional regulatory network (Tnet) [6]. The other two, Mouse GRN and Human GRN, are manually curated databases (TRRUST v2) [7]. Data contain the list of links between transcription factors (TF) and corresponding target genes (TG). All edges have been experimentally confirmed earlier.

First, we studied the GRNs structure using NetworkX 2.8.1 [8] and pandas 1.4.2 Python 3.10.1 libraries [9].

Yeast GRN has 4 441 genes with 12 873 interactions. Of these, 157 genes are TF, and 4 410 are targets. The average number of interactions for nodes is 2.8987. Mouse GRN has 2 456 genes with 7 057 interactions. Of these, 827 genes are TF, and 2 092 are targets. The average number of interactions for nodes is 2.6425. Human GRN has 2 862 genes with 8 427 interactions. Of these, 795 genes are TF, and 2 492 are targets. The average number of interactions for nodes is 2.9444. We used three methods to obtain age classes: protein age classes [2], GenOrigin database [10] and calculated using a phylostratigraphy approach [3].

Protein age classes [2] were translated into gene age classes using protein-gene name matching from the YeastGenome [11] and UNIPROT [12] databases. Interaction maps of TF and targets and TG/TF heatmaps were built for each GRN. Finally, the "difference" of ages in relationships was calculated. The number is the difference between the ages; the smaller, the closer the ages of the interacting genes.

We used the gene ages from the GenOrigin [10] database to calculate the same parameters as for protein classes for Yeast GRN parameters. We used the GenOrigin phylogenetic tree to convert a numerical age into an age class.

We used a phylostratigraphy approach [2] to determine the age of yeast genes in GRN. The iTOL tree [13] phylogeny was used in the analysis to truncate the swiss DB. We compared 4 184 yeast gene sequences by BLAST (blastx) against truncated the Swiss-prot [14] database (94 268 sequences, (10-3 E-value cutoff).

We tested the possibility of randomly obtaining the derived age class ratios in the gene regulation network. We randomly reassigned age classes to 1000 yeast, mouse, and human GRNs to do this. The percentage of each "age" interaction distance for each network was calculated. For each resulting age distance distribution, the standard deviation was counted.

The workflow is represented in the .ipynb files and available in the GitHub repository https://github.com/Freddsle/age_patterns.

Results and Discussion

After translation and mapping protein age classes to GRNs, age was determined for 3 437 (77.4%) genes in Yeast GRN, for 2 287 - (93.1%) in Mouse GRN, and 2 855 (99.8%) - in Human GRN. For the genes, 8 age classes were identified for each GRN. Cellular_organisms, Euk+Bac, Euk_Archaea, Eukaryota, Opisthokonta classes were found in all three networks. Dikarya, Ascomycota, Saccharomyceta classes present in Yeast GRN, and Eumetazoa, Mammalia, Vertebrata in Mouse and Human GRNs.

The proportion of the 'Eumetazoa->Eumetazoa' and 'Eumetazoa->Vertebrata' interactions are the largest among all interactions for mouse and human GRNs (each is more than 10%). On average one TF controls more targets (maximum up to 25) in the yeast network than in mouse (up to 6) and human GRN (up to 8). For yeast GRN, younger nodes have more edges to different age nodes in the network than older nodes. For mouse and human GRNs, the differences are less noticeable. There is no such drop in the number of connections with increasing age.

Human and mouse GRNs have demonstrated a tendency for genes from similar age groups to interact more with each other than with more "distant" age groups. For the yeast GRN, this does not seem to be the case.

The gene ages calculated from the protein ages gave different results for human and mouse, and yeast GRNs. Therefore, we decided to use the gene ages from the GenOrigin. After mapping, we determined the age class for 4 184 genes (94.2%) in Yeast GRN.

TF of the 'Dikarya' age class control fewer targets than other TF classes; there are less than six targets per 'Dikarya' TF. Also, targets of 'Dikarya' and 'Opisthokonta' classes are controlled by m0000ore TF than other target classes. There are less than 5 'Opisthokonta' targets per TF. For 'Dikarya' TF and 'Opisthokonta' targets, the proportion of links among all links in the network is minimal for any edges (less than 0.3% for any combination).

Using gene ages from the GenOrigin, there is no significant predominance of interactions between similar age classes in the yeast network. Edges with age distances 0 ("same age") and 1 ("close age") account for less than 35% of all edges.

When using phylostratigraphy, the fraction of "same age" interactions (distance between ages is 0) has increased. However, this observation may be caused by the truncated tree, in which all age classes older than eukaryotes also received the label eukaryotes. Also, even though the 'Opisthokonta' class was sufficiently represented in the truncated Swiss database, the number of targets of this age class turned out to be less than expected. Therefore, we plan to blast GRNs genes to a fine-grained tree with a more uniform representation of nodes across gene classes.

Was it possible to obtain preferences in the interaction in a random network? We determined interaction preference only for certain age distances (distances are 2, 7 or 8) using the method with randomly assigned classes in Yeast GRN.

Conclusion

Unfortunately, we cannot confidently say that our hypothesis about gene interaction preference in GRN has been confirmed. None of the three methods used to obtain gene ages showed that interactions of "same" and "close" age are dominated in yeast GRN. There are no significant differences compared to the model where the

age categories are randomly assigned. We need a more correctly formulated null hypothesis (a method for obtaining a random network) or a more correct phylogenetic resolution (a fine-grained tree with a more uniform representation of nodes across gene classes).

References

1. Davidson, Eric H. "Emerging properties of animal gene regulatory networks." Nature 468.7326 (2010): 911-920.

2. Liebeskind, B.J., McWhite, C.D., and Marcotte, E.M. "Towards consensus gene ages." Genome biology and evolution 8.6 (2016): 1812-1823.

3. Domazet-Lošo, T., Brajković J., and Tautz, D. "A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages." Trends in Genetics 23.11 (2007): 533-539.

4. Chen, C-Y., et al. "Dissecting the human protein-protein interaction network via phylogenetic decomposition." Scientific reports 4.1 (2014): 1-10.

5. Capra, J.A., Pollard, K.S., Singh, M.. "Novel genes exhibit distinct patterns of function acquisition and network integration." Genome biology 11.12 (2010): 1-16.

6. Balaji, S., et al. "Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast." Journal of molecular biology 360.1 (2006): 213-227.

7. Han, H., et al. "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions." Nucleic acids research 46.D1 (2018): D380-D386.

8. Hagberg, Aric, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. No. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

9. Pandas Python Library. Link: https://pandas.pydata.org/.

10. Tong, Y.-B., et al. "GenOrigin: A comprehensive protein-coding gene origination database on the evolutionary timescale of life." Journal of Genetics and Genomics (2021).

11. The Saccharomyces Genome Database. Link: https://www.yeastgenome.org/.

12. The UniProt Consortium. “UniProt: the universal protein knowledgebase in 2021”. Nucleic Acids Res. 49 (2021): D1.

13. Letunic, I., Bork, P.. "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation." Nucleic acids research 49.W1 (2021): W293-W296.

14. Bairoch, Amos, and Rolf Apweiler. "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000." Nucleic acids research 28.1 (2000): 45-48.

УДК 577.22

A transcriptome assembly from fragments of the annelids *Pygospio elegans* (*Spionidae*, *Annelida*) and *Arenicola marina* (*Arenicolidae*, *Annelida*)

A. Chen[1], V. Starunov[1,2], Z. Starunova[2] and E. Novikova[1,2]

[1] *Saint-Petersburg State University, Russia*
[2] *Zoological Institute RAS, Saint-Peterburg, Russia*

Introduction

Reparative regeneration, i.e. the ability to regrow lost body parts or structures after natural ablation or amputation is considered to be an ancestral feature of all metazoans [2]. Regenerative capacities have been lost multiple times during evolution in variable groups of animals and the reasons for this loss are mostly unknown [1]. One of the important steps of reparative regeneration is to correctly integrate the regrowing body part or structure with the existing tissue [9]. This can be reached by establishing and maintaining the positional information in the body which can be reorganized after injury in consistency with the new proportions and size of the body. Positional information is the set of molecular signals which is gradientally spread in the body thus each cell of the body responds to a certain level of a certain molecule [11].

The gradients like this were described for a number of animals [5,7,8,10]. In planarians, the posterior-anterior gradient of Wnt-signaling is maintained in the adult body and its insufficiency correlates with the absence of head regeneration [5], [8]. In *Danio rerio* a number of factors establish the gradiental expression along the fin rays and probably maintain its ability to regenerate through the lifespan [7]. Similarly, the gradiental expression of a number of Hox genes was demonstrated in the juvenile annelid *Alitta virens* (Nereididae) and those gradients are reorganized after the “tail” amputation and its further restoration [8]. Thus, the ability to establish and maintain positional information in the juvenile or adult body seems to correlate with good regeneration capacities. Annelids, or ring worms, possess remarkable ability to restore anterior and posterior parts of the body, although this ability varies a lot in different annelid families [4]. For example, the spinoid *Pygospio elegans* can easily restore head and tail parts after amputation sometimes form a single segment [6]. On the contrary, the arenicolid *Arenicola marina* can only heal the wound and compensate for the loss of the size by hypertrophic growing of the remaining segments [3]. We suggest that the system of maintaining molecular gradients can exist in *P. elegans* body and is probably absent in *A. marina*. To test this hypothesis, we cut the worms into 12 parts, isolated total RNA and created the transcriptome for each body piece. Here, in this paper we describe the processing of the raw reads and further transcriptome assembly.

Methods

Paired-end Illumina reads were preprocessed before analyzing the data in the following way: quality control was performed using FastQC [12], and sequencing errors were corrected via Karect [13]. Then, low-quality and adapter sequences were clipped with Trimmomatic (with parameters sliding window:5:20, leading:25, trailing:25, minlen:25) [14].

The prepared reads were assembled with Trinity [15]. We then checked statistics of the assembly with rnaQUAST in order to perform the quality assessment [16]. Trinity *de novo* assembly has artificial redundancy due to the use of De Bruijn graphs. To get rid of that we used CD-HIT that clusters similar sequences into clusters [17].

Next, we predicted the location of ribosomal RNA genes using Barrnap [18]. Decontamination was removed via MCSC [19]. This method is based on a hierarchical clustering algorithm and uses the UniRef90 database to identify contaminant clusters. We ran MCSC setting taxon to keep as '*Annelida*'.

After having clustered and decontaminated data, we analyzed gene expression. For this purpose, we used Salmon [20] to produce transcript-level quantification estimates from our data. We were then able to use a library tximport for R to summarize expression to genes [21]. Then, we identified candidate coding regions within transcript sequence via TransDecoder [22]. We first extracted the long open reading frames and next searched the peptides for protein domains using Pfam and Hmmer3 [23].

Results

FastQC revealed failures in following sections: per base sequence content, GC content, sequence duplication levels, overrepresented sequences, adapter content. Some of them are just a feature of RNA data, but low quality and adapter sequences needed to be clipped. After running Karect and Trimmomatic approximately 10% of reads were dropped. rnaQUAST showed that Trinity assembly of *Pygospio elegans* had 910828 contigs, while assembly of *Arenicola marina* had 355266 contigs. Moreover, the report indicated that *Arenicola marina* assembly had higher average contig length and better Nx statistics.

After getting rid of artificial redundancy with CD-HIT we obtained 615358 contigs for *Pygospio elegans* and 271748 clusters of contigs for *Arenicola marina*. Next step of the assembly post-processing was decontamination. We used Barrnap and then searched the results against the NCBI database. This revealed contamination with *Selenidium pygospionis*. MCSC was used to get rid of the contamination and as a result we obtained 423597 and 185289 sequences for *Pygospio elegans* and *Arenicola marina* respectively.

Based on the results of Salmon and TransDecoder, we were ready to get protein coding genes with significant expression. Out of all genes we selected those that have >1 TPM expression. Finally, we chose genes that encode proteins longer than 100 amino acids. Two sets of protein-coding genes selected with significant expression:

54315 (*Pygospio elegans*) and 33530 (*Arenicola marina*).

Discussion

The task of correctly identifying contamination and removing such data from future analysis can be solved using one of the two approaches: by direct comparison sequences with the database and removing those that have the greatest similarity or through the classification of sequences using distinctive features. Comparison with the database has an important limitation. The database itself can miss data for the species that might interest us. That is, some sequences (large or small - depending on the studied group or species) will be identified incorrectly, or will not have any annotation at all. For this reason, MCSC algorithm was chosen. According to the publication [12] describing the program, the algorithm analyzes properties of the sequences, isolating some specific patterns and classifying all sequences in a transcriptome based on similarities and differences in their patterns. This approach is justifiable, but does not always guarantee that the output is completely decontaminated. However, it is the best solution we were able to find.

We should mention that *Arenicola marina* assembly has significantly less contigs. However, statistics like average contig length and Nx are better. Moreover, the resulting gene sets that were obtained also vary in number. There are almost twice as many obtained *Pygospio elegans* genes as *Arenicola marina* genes. This might imply that such a difference comes from the initial sequencing data we had at hand. In order to solve this issue, more samples should be considered. There are two more samples of considered annelids available with sequencing data. The same process can be applied in order to validate the results. Prepared data, *Pygospio elagans* and *Arenicola marina* assemblies and sets of genes, can be further analyzed in order to determine gene-candidates responsible for positional information concept.

The research was supported by RSF grant # 21-14-00304.

References

1. Bely AE. Evolutionary Loss of Animal Regeneration: Pattern and Process. Integrative and Comparative Biology. 2010. 50(4): 515–527.

2. Bely AE, Nyberg KG. Evolution of animal regeneration: re-emergence of a field. Trends Ecol Evol. 2010; 25(3):161-170.

3. Berrill N. J. Regeneraton and budding in worms. Biological Reviews. 1952. 27:401–438.

4. Kostyuchenko RP, Kozin VV. Comparative Aspects of Annelid Regeneration: Towards Understanding the Mechanisms of Regeneration. Genes (Basel). 2021. 12(8):1148.

5. Liu SY, Selck C, Friedrich B, Lutz R, Vila-Farré M, Dahl A, Brandl H, Lakshmanaperumal N, Henry I, Rink JC. Reactivating head regrowth in a

regeneration-deficient planarian species. Nature. 2013. 500(7460):81-4.

6. Malikova I.G., Plusch T.A. [Morphogenetic processes during the regeneration of the polychaete Pygospio elegans from fragments of the body]. 1980. Ms. Deposited in VINITI. 1250–80:18. [In Russian]

7. Nachtrab G, Kikuchi K, Tornini VA, Poss KD. Transcriptional components of anteroposterior positional information during zebrafish fin regeneration. Development. 2013. (18):3754-64.

8. Novikova EL, Bakalenko NI, Nesterenko AY, Kulakova MA. Expression of Hox genes during regeneration of nereid polychaete Alitta (Nereis) virens (Annelida, Lophotrochozoa). Evodevo. 2013. 4(1):14.

9. Oderberg IM, Li DJ, Scimone ML, Gaviño MA, Reddien PW. Landmarks in Existing Tissue at Wounds Are Utilized to Generate Pattern in Regenerating Tissue. Curr Biol. 2017; 27(5):733-742.

10. Sikes JM, Newmark PA. Restoration of anterior regeneration in a planarian with limited regenerative ability. Nature. 2013. 500(7460):77-80.

11. Wolpert L. One hundred years of positional information. Trends in Genetics. 1996 12: 359–364.

12. FastQC. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

13. Karect. https://github.com/aminallam/karect

14. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

15. Trinity. https://github.com/trinityrnaseq/trinityrnaseq/wiki

16. Bushmanova, E., Antipov, D., Lapidus, A., Suvorov, V. and Prjibelski, A.D., 2016. rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. Bioinformatics, 32(14), pp.2210-2212.

17. Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, Weizhong Li, CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics*, Volume 28, Issue 23, 1 December 2012, Pages 3150–3152.

18. Barrnap. https://github.com/tseemann/barrnap

19. Joël Lafond-Lapalme, Marc-Olivier Duceppe, Shengrui Wang, Peter Moffett, Benjamin Mimee, A new method for decontamination of *de novo* transcriptomes using a hierarchical clustering algorithm, *Bioinformatics*, Volume 33, Issue 9, 1 May 2017, Pages 1293–1300.

20. Salmon. https://github.com/COMBINE-lab/salmon

21. Soneson C, Love MI, Robinson MD (2015). “Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences.”

F1000Research, 4.

22. TransDecoder. https://github.com/TransDecoder/TransDecoder/wiki

23. hmmer3. https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan

УДК 575.85

Application of machine learning methods to approximate demographic history parameters from allele frequency spectrum

E. Gorelkina

Bauman Moscow State Technical University, 105005, Moscow, 2nd Baumanskaya str., 5, p. 1

Introduction

Demographic history is the reconstructed record of the population in the past. The demographic history of populations is determined by a number of parameters such as: the time of split, the rate of migration, the size of subpopulations after split, and others.

The allele frequency spectrum is a statistic of demographic history data: it is the distribution of the allele frequencies of a given set of loci (often SNPs) in a population or sample. For two populations, such statistics can be represented by a two-dimensional tensor.

Currently existing optimization methods that allow for optimizing the parameters of demographic history by its allele frequency spectrum have a substantial running time, although one of the approaches [1] uses a genetic algorithm that allows you to accelerate optimization. We would like to get predictions in a faster and more accessible way using a machine learning model.

The goal of this exploration is to study the effectiveness of machine learning methods for quickly predicting the parameters of a demographic history, understand how to validate such a model, and draw conclusions about the effectiveness of the metrics used.

Materials and methods

The 2_DivMig_5_Sim model [2] was chosen as a demographic history model for study. Then, by changing the parameters, allele-frequency spectrum datasets were generated for training, validation, and testing of the machine learning model. The vector of demographic history parameters was used as the target function. As a consequence, we needed a multi-output regression model.

A random forest [3] was taken as a machine learning model, since this model is quite simple on the one hand and, on the other hand, is a good regressor, since it can build quite complex regression curves and at the same time is an ensemble model, which increases the accuracy of predictions, unlike, for example, a decision tree.

For the implementation of the multi-output regression model we used two approaches. The first is to predict the parameters independent of each other (model I). The second is to predict parameters correlated to each other, while the order of dependence is determined by the order of parameters in the parameter vector (model II).

We also used metrics for validation of our machine learning model. The first metric we used was the coefficient of determination (R^2 metric) [4]. The second metric is the random search metric. To get the averaged value, it is necessary to perform this algorithm for each spectrum from the test sample; the predicted spectrum is the spectrum generated by the parameters predicted using the ML model. Averaging over all test spectrums, we get the random search metric.

Results

We have developed a pipeline for conducting such research, which includes generating datasets, storing them as files, preprocessing data, training and validating models, testing models and using a random search algorithm. A small Python script was also developed that allows you to make predictions based on pre-trained models.

All developed scripts and pipelines can be found here: https://github.com/lisosoma/ML-for-demographic-inference.

Discussion

As a result of our exploration, we found that a machine learning model such as a random forest copes well enough with predictions of demographic history parameters on the allele frequency spectrum. In general, it can be concluded that the studied approach allows you to quickly assess the parameters of demographic history, while losing accuracy.

For the two datasets we generated, we got quite good metrics: for all models, R^2 shows results of more than 0.89. As for random search, its metric for all models exceeds 850. This means that we need at least 850 random points to achieve the same likelihood (with a certain accuracy) as a random forest gives. It is worth noting that the more such random points are required, the better the machine learning model works for these kinds of predictions.

It is worth noting that there is no strong difference in metrics between the two approaches to predictions: with independent parameters and correlated ones. Based on this, it can be assumed that the parameters of demographic history are weakly dependent on each other.

For further research, it is proposed to try other machine learning models, in particular, a convolutional neural network or generative-adversarial models. It is also worth trying to use other demographic history models. As for the software implementation, one can develop a class and use pre-trained machine learning models

for each demographic history model in it. This will automate the forecasting process. It is also worth suggesting some normalization of the random search metric. Also, a metric based on random search needs a lot of computing costs, so it is better to use the GPU to perform this pipeline.

References

1. Ekaterina Noskova, Vladimir Ulyantsev, Klaus-Peter Koepfli, Stephen J O'Brien, Pavel Dobrynin, GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data, GigaScience, Volume 9, Issue 3, March 2020, giaa005, https://doi.org/10.1093/gigascience/giaa005

2. Ekaterina Noskova's repository where the demographic history model was taken from: https://github.com/noscode/demographic_inference_data/tree/master/2_DivMig_5_Sim

3. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

4. Glantz, Stanton A.; Slinker, B. K. (1990). Primer of Applied Regression and Analysis of Variance. McGraw-Hill. ISBN 978-0-07-023407-9.

УДК 577.22

Search for homologs of egg-cell specific genes, study of their expression patterns and regulatory elements for the creation of effective constructs for genetic engineering

E. Grigoreva[1], A. Toidze[2], M. Logacheva[3], A. Kasianov[4]

[1] *Institute of Cytology and Genetics, Siberian branch, Russian Academy of Sciences, Prospekt Lavrentyeva 10, Novosibirsk, Russia, 630090*
[2] *Faculty of Chemistry and Biochemistry, Ruhr-Universität Bochum, Universitätsstraße 150, 44780 Bochum, Germany*
[3] *Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, bld. 1, Moscow, Russia, 121205*
[4] *Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoy Karetny per. 19, bld.1, Moscow, Russia, 127051*

Genome editing using CRISPR/Cas9 enables the study and the production of plants with improved traits for agriculture. The CRISPR/Cas9 system consists of two components: the Cas9 protein and the guide RNA (gRNA). The gRNA is a small 20 nt long RNA that provides sequence specificity, forms a complex with Cas9 and guides the Cas9 protein to its target sequences. Cas9 is an RNA-dependent DNA endonuclease that produces double-stranded breaks at the target sites. The DNA repair systems in host cells are typically induced after cleavage by Cas9. Constitutive promoters such as the 35S promoter or promoters of housekeeping genes have been used for genetic engineering in plants, as they have high levels of expression in all cell types. However, transgenic lines with these promoters have been mainly mosaic in the first generation. Usage of egg cell-specific promoters has been shown to enable the creation of non-mosaic T1 mutants for multiple target genes with high efficiency [1]. EC1.1 and EC1.2 are Arabidopsis thaliana genes from the egg cell-specific gene family that are specifically and highly expressed in egg cells. Assuming that homologous genes have similar functions, we supposed that EC homologs could have similar expression patterns in other plants and using their promoters could improve genome editing in corresponding plants. Thus, the aim of our project is to find functional analogs of EC genes in different crops and model plants and explore their expression patterns and regulatory elements.

In this study we used genomes, annotations and amino acid sequences of 53 plant species from different public databases (Plant Ensemble, PLAZA, MBKBASE and Phytozome). We used the Orthofinder v.2.5.4. to find EC1 gene orthologs. In Orthofinder, EC1.1 and EC1.2 genes were grouped into one orthogroup. Protein sequences of all genes that were in the same orthogroup as EC1 genes (201 genes) were taken for further phylogenetic analysis. Alignment obtained by Clustal Omega v1.2.3 was used to construct the phylogenetic tree using IQ-TREE v2.0.3a by maximum likelihood method using ultrafast bootstrap approximation. Amborella

trichopoda was chosen as the outgroup. Two clades with very high bootstrap support formed on the tree, which roughly correspond to the EC1.1 and EC1.2 gene families. Inside the clades genes are grouped according to species phylogeny. These clades contain the majority of genes of both dicots and monocots. This could imply that the duplication leading to the emergence of EC1.1 and EC1.2 occurred in the early stages of the evolution of flowering plants, even before divergence of dicots and monocots. The structure of the tree within each of the clades is consistent with the phylogeny of flowering plants. However, genes from monocots are present in only one clade, EC1.1. This suggests that the common ancestor of the monocots lost one of the paralogs corresponding to EC1.2 The outside group probably contains genes that are not EC1.1 or EC1.2 orthologs (and are there due to, e.g., long branch attraction).

Based on Orthofinder results, some species had several orthologs of EC1 genes. In order to determine the genes that most probably have functions most similar to Arabidopsis egg-cell specific genes, the expression patterns of the orthologs were examined. Due to the absence or bad quality of openly accessible RNA-seq data, only 20 species and 53 orthologs of these organisms were taken for further analysis. According to their expression profiles, genes were divided into three groups. The first group contained 23 genes with expression profiles similar to the EC1 gene family in A. thaliana. They were highly and specifically expressed in female reproductive organs. The second group contained 29 genes that have expression in female reproductive organs as well as other plant tissues. Here, the expression in the female reproductive organs was not the highest. Finally, the third group contained genes that have no expression in female generative organs, only in vegetative parts of plants and contains 5 genes.

For each group, we searched for characteristic motif sequences in the upstream regions of the genes. From the Jaspar 2022 database we took all known motif sequences specific for plants (656 motifs) and used the FIMO v5.4.1 to search for these motifs in the 500 bp upstream of found orthologs. Using custom python scripts, we generated heatmaps to search for patterns in the occurrences of the motifs. We also looked at motifs that are present in more than 50% genes in groups 1 (only female reproductive organs) and groups 2 (female reproductive organs and other plant tissues). However, no specific motif was found that universally accounts for the specific expression in female reproductive organs. Possibly, it is only enabled by a specific combination of different motifs and for each gene this combination is unique.

References

1. Wang, ZP., Xing, HL., Dong, L. et al. Egg cell-specific promoter-controlled CRISPR/Cas9 efficiently generates homozygous mutants for multiple target genes in Arabidopsis in a single generation // Genome Biol — V16, №144 — 2015 — pp. 1-12.

УДК 575.85

Molecular mechanisms behind the life cycle evolution and speciation in hydroids of the Arctic region

P. Guro[1], L. Danilov[2,3], S. Kremnyov[4,5]

[1] *All-Russia Research Institute for Agricultural Microbiology, Sh. Podbelsky 3, Pushkin-8, 196608, St. Petersburg, Russia*
[2] *Department of Genetics and Biotechnology, St. Petersburg State University, Universitetskaya emb. 7/9, 199034, St. Petersburg, Russia*
[3] *Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia*
[4] *Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia*
[5] *Koltzov Institute of Developmental Biology of Russian Academy of Sciences, 26 Vavilov Street, Moscow, 119334, Russia*

Hydrozoans are a group of cnidarians that are noted for their complexity and diversity in life cycles. In many hydrozoan species, the life cycle consists of a free-living planula larva that transforms into a primary polyp. The primary polyp buds other polyps to produce a benthic colonial stage. Upon reproductive maturity, the polyps bud pelagic medusae that ultimately form gametes and spawn in the water column [1]. Within hydrozoans there exists an extraordinary variation in this life cycle that is reflected in a wide range of diversity of polyp, colony, and medusa morphologies, as well as complete loss or reduction of the polyp or medusa stage in some species. The molecular mechanisms of speciation in hydroids have never been studied, neither has the relationship between the evolution of the life cycle and speciation ever been considered. The hydroid *Sarsia lovenii* from the White Sea was chosen as the object. Recently, in *S. lovenii*, breeding season polymorphism has been found to be associated with life cycle polymorphism [2]. Colonies of the first morph produce normally developed free-floating medusas, while colonies of the second morph produce attached gonophores - medusoids. The morphs identified represent phenological populations: in the example of *S. lovenii*, we can observe the initial stage of sympatric speciation. Thus, due to the aforementioned features of the object we have chosen, we can study the molecular mechanisms of speciation associated with the divergence of populations in breeding time and associated with the evolution of the life cycle.

Transcripts of two morphotypes - medusa and medusoids were assembled *de novo* and annotated. Assembled transcriptome size was 50 Mbp and 58 Mbp, respectively. Next, analysis of differential gene expression was performed. We used 12 libraries with sequences from big and small medusa and medusoids buds samples, respectively (100 nt, single end) and analyzed the differential expression of the 50 most up and down regulated genes. Since the medusa stage is considered to be the ancestral state of hydrozoan, we focused on genes that could be a marker of the difference between medusa and medusoid stages. Thus, we identified 5 homeobox genes for the related genus *Clytia*, that are assumed to be specific to the medusas

stage - *Tlx, Pdx, DRGX, CnoxA, Cnox4*. Then we searched for selected genes in medusa and medusoid buds samples, respectively. It was found that all five genes were expressed in all medusa buds samples and only two genes were expressed in medusoid buds samples - *CnoxA* and *DRGX*. We found in the new preprint article authors suggest that the *Tlx* homeobox genes play a key role in medusa development and the loss of this gene is probably related to the loss of the medusa life cycle stage [3]. Expression of *CnoxA, DRGX* genes in samples that have lost medusa stage and the absence of *Tlx* gene expression may support the assumption that we are observing the initial stage of sympatric speciation. Data that we obtained can be a source for further studies of mechanisms that are associated with the loss of the medusa stage in the life cycle of hydrozoa.

References

1. Paulyn Cartwright, Annalise M. Nawrocki, Character Evolution in Hydrozoa (phylum Cnidaria), Integrative and Comparative Biology, Volume 50, Issue 3, September 2010, Pages 456–472

2. Prudkovsky, Andrey A., Irina A. Ekimova, and Tatiana V. Neretina. "A case of nascent speciation: unique polymorphism of gonophores within hydrozoan Sarsia lovenii." Scientific reports 9.1 (2019): 1-13

3. Travert, Matthew Kevin, et al. "Coevolution of the Tlx homeobox gene with medusa development (Cnidaria: Medusozoa)." bioRxiv (2022)

УДК 004.94, 579.25

Studying complex structural variations in cancer using long reads

O. Kalinichenko [1], M. Kolmogorov [2]

[1] *Moscow Institute of Physics and Technology, 1 "A" Kerchenskaya st., 117303, Moscow, Russian Federation*

[2] *National Institutes of Health / National Cancer Institute, 9609 Medical Center Drive, Building 9609 MSC 9760, Bethesda, Maryland 20892-9760, USA*

Introduction

Cancer has been known and feared for more than two thousand years, but mechanisms of its genesis seemed mysterious for most of the time. Now it is known that cancer is caused by genetic changes. There are numerous mutations in tumor cells including single nucleotide polymorphisms (SNPs), insertions, deletions, inversions, translocations, and chromosomal aberrations [10]. Small mutations have been studied thoroughly using short read sequencing. However, big structural variations are hard to be determined using only short reads due to read mapping ambiguities [9, 11]. Recent advances in long read sequencing present an opportunity to solve this problem. There are still many problems to cope with, such as the heterogeneity of tumor cells and loss of information about haplotypes present in a read (one read may consist of parts from different haplotypes if there is a breakpoint inside it).

In this study, we use long read data to determine complex rearrangements in cancer, including information about haplotypes that form "new" cancer chromosomes. In particular, we focus on automatically determining possible breakage-fusion-bridge events and investigate these regions more thoroughly. We combine finding structural variations breakpoints and analyzing read coverage change (corresponding to copy number change) to solve this problem.

Methods and materials

Data

We used an alignment of Oxford Nanopore reads to GRCh38 human reference in bam format. Each read was already phased according to its primary alignment (this information was stored in the HP tag).

Methods

For finding breakpoints, we used Sniffles [1, 2] and a tool adapted from the source code of HapDup [3, 4, 5]. Sniffles provided a less accurate result, so we focused on the HapDup result.

After that, we developed a Python3 script for visualizing read coverage by haplotype and breakpoints. We also developed a script for automatically determining possible breakage-fusion-bridge events by analyzing coverage profile and breakpoints. We analyzed these locations and performed local assembly with Flye

assembler (version 2.9) [6, 7] around the breakpoints to support or contradict our hypothesis. We chose Flye assembler because it was specially designed for ONT reads.

The alignments were then examined in the IGV genome browser (version 2.11).

Results

We developed a Python3 script for visualizing read coverage by haplotype and breakpoints. It works on a BAM file and produces pictures in PNG format. It creates visualizations of every chromosome and also a zoom-in of all possible breakage-fusion bridge events.

The script can be found here: https://github.com/madshuttlecock/structural-variations. The script can be used as a console script (script.py) and also all functions can be imported in Python3 separately (just import counter.py).

We also created a script for automatically determining possible breakage-fusion-bridge events. There are possible events on chromosomes 3 and 15.

Discussion

We found possible breakage-fusion-bridge events on two chromosomes. However, these variations usually happen on the ends of chromosomes, and in our case they are observed closer to the middle. We suppose that it is due to the loss of a part of the chromosome. We are planning to verify this hypothesis. Another possible explanation is the formation of an extrachromosomal DNA (ecDNA), but we believe it to be less likely.

We are currently working on combining all our tools to automatically determine all structural variations and the full chromosomal structure in cancer cells.

References

1. Sedlazeck, F.J., Rescheneder, P., Smolka, M. et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15, 461–468 (2018). https://doi.org/10.1038/s41592-018-0001-7

2. https://github.com/fritzsedlazeck/Sniffles

3. Kishwar Shafin, Trevor Pesout, Pi-Chuan Chang, Maria Nattestad, Alexey Kolesnikov, Sidharth Goel, Gunjan Baid et al. "Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks." bioRxiv (2021). doi:10.1101/2021.03.04.433952

4. Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin and Pavel Pevzner, "Assembly of Long Error-Prone Reads Using Repeat Graphs", Nature Biotechnology, 2019 doi:10.1038/s41587-019-0072-8

5. https://github.com/fenderglass/hapdup

6. Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin and Pavel Pevzner, "Assembly of Long Error-Prone Reads Using Repeat Graphs", Nature Biotechnology, 2019 doi:10.1038/s41587-019-0072-8

7. https://github.com/fenderglass/Flye

8. Twelve years of SAMtools and BCFtools, Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li, GigaScience, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008

9. A multi-platform reference for somatic structural variation detection, Jose Espejo Valle-Inclan, Nicolle J.M. Besselink, Ewart de Bruijn, Daniel L. Cameron, Jana Ebler, Joachim Kutzera, Stef van Lieshout, Tobias Marschall, Marcel Nelen, Andy Wing Chun Pang, Peter Priestley, Ivo Renkens, Margaretha G.M. Roemer, Markus J. van Roosmalen, Aaron M. Wenger, Bauke Ylstra, Remond J.A. Fijneman, Wigard P. Kloosterman, Edwin Cuppen, doi: https://doi.org/10.1101/2020.10.15.340497

10. Lessons from the Cancer Genome, Levi A. Garraway, and Eric S. Lander, http://dx.doi.org/10.1016/j.cell.2013.03.002

11. Reconstruction of clone- and haplotype-specific cancer genome karyotypes from bulk tumor samples, Sergey Aganezov, and Benjamin J. Raphael

УДК 577.22

Analysis of differential expression of genes involved in NO-signaling in synucleinopathies

A. Kapitonova[1], A. Livanova[2], S. Bondarev[2]

[1] *Lomonosov Moscow State University, 119991, 1 Leninskiye gori, Moscow, Russia*
[2] *Saint-Petersburg State University, 199034, 7-9 Universitetskaya Embankment, Saint-Petersburg, Russia*

Introduction

Synucleinopathies are neurodegenerative diseases that include Parkinson's disease, multiple system atrophy, and dementia with Lewy bodies (Coon and Singer, 2020). In the course of pathogenesis, protein aggregates, in particular, alpha-synuclein, are formed in neurons of some subcortical nuclei of the brain, which leads to loss of control of voluntary and involuntary movements in patients. According to bioinformatics predictions, the adapter protein of nitric oxide synthase 1 (NOS1AP) is also capable of forming protein aggregates in neurons (unpublished data). NOS1AP is a cytosolic protein that binds to neuronal nitric oxide synthase (NOS1), which is responsible for nitric oxide production in neurons. Besides NOS1, NOS1AP interacts with a number of other proteins and could be involved in various processes: Hippo signaling pathway (controls cell proliferation and differentiation) (Clattenburg et al., 2015), dendrite development (Candemir et al., 2016; Carrel et al., 2009), control of circadian rhythms, iron homeostasis in neurons, and NMDA/NO-mediated neurotoxicity (Wang et al., 2016). NOS1AP also participates in the nNOS-p38MAPK signaling pathway involved in excitotoxicity (toxicity of excitatory neurotransmitters such as glutamate), a phenomenon described for many neurodegenerative processes (Li et al., 2013). NOS1AP is known to be implicated in neuropsychiatric disorders such as post-traumatic stress disorder (Lawford et al., 2013), bipolar disorder (Freudenberg et al., 2015), and schizophrenia (Brzustowicz et al., 2004; Miranda et al., 2006; Zheng et al., 2005). An increase in NOS1AP expression occurs in response to spinal cord injury in rats. It was assumed that this fact might be related to the subsequent death of neurons. At the same time, accumulations of this protein are detected on histological sections of nerve fibers at the site of injury (Cheng et al., 2008). This result may confirm the formation of NOS1AP aggregates in vivo, as well as their neurotoxic effect. In addition, there is evidence allowing to link NOS1AP with neurodegeneration in the development of Huntington's and Alzheimer's diseases (Wang et al., 2016). Furthermore, it has recently been shown that an increase in NOS1AP induces tau protein aggregation as well as neurodegeneration (Hashimoto et al., 2019). Finally, the ability of this protein

to directly interact with alpha-synuclein allowed us to assume that NO signaling could be involved in the pathogenesis of synucleinopathies. The aim of this project was to evaluate changes in expression level of the NOS1AP gene and other NO-signaling genes in brain samples from patients with synucleinopathies.

Materials and methods

Four sets of RNA sequencing data of patients with synucleinopathies were selected from the open SRA database: i) SRP058181, Brodmann area prefrontal cortex, Parkinson's disease (n=29), control (n=42); ii) SRP148970, substantia nigra, ventral tegmental area (VTA), Parkinson's disease (n=5), control (n=18), iii) SRP215213, putamen, multiple system atrophy (n=10), control (n=12), iv) SRP324001, anterior cingulate cortex, dementia with Lewy bodies (n=7), Parkinson's disease (n=7), Parkinson's disease with dementia (n=7), control (n=7). Available metadata of samples, such as gender and age of patients, were extracted from SRA database and corresponding publications. The quality of raw reads was assessed in FastQC v0.11.5 (Andrews, 2010) and summary reports were created in MultiQC v1.12 (Ewels et al., 2016). Alignment against the GRCh38 human genome was performed with STAR v2.7.10a (Dobin et al., 2013) using GeneCounts option. Principal component analysis (PCA) with rlog transformed counts was performed to confirm the clustarization of groups. Two protocols were used to deal with technical replicates, where ones were summarized or averaged, neither of which affecting results obtained in further analysis. Variant calling in genes associated with synucleinopathies (SNCA, LRRK2, GBA, PRKN) was performed via Samtools htslib 1.10.2-3 (Danecek et al., 2021). Clinical significance of mutations was assessed with Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016).

DESeq2 library (Love et al., 2014) was chosen to perform differential expression analysis with Log2FoldChange (lfc) correction using apeglm method (Zhu et al., 2019). The following thresholds for significant differential expression were chosen: s-value < 0.005, |lfc| > 1. Lists of genes with significantly changed expression were uploaded to Gene Ontology, gsea and kobas databases to reveal main signaling pathways upregulated and downregulated in synucleinopathies. The EnchancedVolcano (Blighe et al., 2019) and VennDiagram R packages were used to draw volcano plots and Venn's diagrams, respectively.

To reveal expression changes of certain genes we used the table with normalized counts corrected for library size. Normalized counts for NOS1AP were compared in controls and patients independently within four datasets (Mann-Whitney test) using GraphPad Prism 8 v8.0.0. To reveal other genes of NO-signaling with changed differential expression, the list of NO related genes was obtained from Gene Ontology by the key word "nitric oxide" and crossed with the lists of genes with significantly changed expression from each of brain tissues.

Results and Discussion

PCA analysis of differential expression data obtained from four RNA seq datasets has revealed two outliers in SRP058181 dataset, which were excluded from the analysis. No clinically significant variants (according to the ClinVar database) were found within variant calling, and stratification of patients by sex and age did not change the pattern of group clustering assessed by the PCA.

Our analysis revealed the following rates of differential expressed genes (DEGs): 0.3% (SRP058181, Parkinson's disease (PD), prefrontal cortex); 1.0% and 1.1% (SRP148970, substantia nigra and ventral tegmental area (VTA), respectively); 0.1%, 0.008%, 0.04% (SRP324001, PD, PD with dementia, and dementia with Lewy body, respectively), 1.8% (SRP215213, multiple system atrophy, putamen). According to GO terms, biological processes significantly altered in analyzed samples, were: i) response to unfolded protein, response to topologically incorrect protein, cellular response to ultraviolet in Parkinson's disease, Brodmann area 9, prefrontal cortex (SRP058181); ii) cytoskeletal protein binding, negative regulation of cell death in Parkinson's disease, substantia nigra; protein containing complex organization, intracellular transport, nitrogen compound transport in Parkinson's disease, VTA (SRP148970); iii) cellular response to copper, cadmium, zinc ions, inflammatory response, negative regulation of transport, specification of symmetry, aerobic respiration, regulation of transcription by RNA polymerase II in multiple system atrophy, putamen (SRP215213). No biological processes were found to be altered in gene enrichment analysis of SRP324001 dataset (anterior cingulate cortex).

Expression of NOS1AP did not differ significantly in brain tissues of patients with synucleinopathies, although a decrease in expression is observed in prefrontal cortex ($p = 0.08$) and substantia nigra ($p = 0.09$) of patients with Parkinson's disease, as well as in the anterior cingulate cortex ($p = 0.142$) in patients with dementia with Lewy bodies. However, among the obtained differentially expressed genes, genes involved in NO signaling were found: i) FOXJ1, LINC01338, KCNE4, HSPA1B_4, HSPA6, HSPA1B_2 in Parkinson's disease, Brodmann area 9, prefrontal cortex (SRP058181); ii) DNM3, SNTG1, JAK2 in Parkinson's disease, substantia nigra (SRP148970); iii) IFNG, TLR2, SPR, CCL2, CD36, AQP1 in multiple system atrophy, putamen (SRP215213) iv) TNFSF12, ADH5 in Parkinson's disease, VTA (SRP324001).

Obtained Venn's diagram showed that only a few common differentially expressed genes were found for different tissue types; patterns of DEGs differed tissue- and disease-specifically.

Thus, the results of our work demonstrate that the expression of some genes involved in NO signaling significantly changes in patients with synucleinopathies. The possible role of these genes and their products in the pathogenesis of these

diseases requires further study. It has been also shown that the differential expression of genes in synucleinopathies is tissue-specific, since an almost unique set of differentially expressed genes was obtained for the transcriptome of each studied brain area. However, it should be pointed out that this result requires further verification using new data with greater sample size.

References

1. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data.

2. Blighe, K., Rana, S., & Lewis, M. (2019). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version, 1(0).

3. Brzustowicz, L. M., Simone, J., Mohseni, P., Hayter, J. E., Hodgkinson, K. A., Chow, E. W., & Bassett, A. S. (2004). Linkage disequilibrium mapping of schizophrenia susceptibility to the CAPON region of chromosome 1q22. American journal of human genetics, 74(5), 1057–1063.

4. Candemir, E., Kollert, L., Weißflog, L., Geis, M., Müller, A., Post, A. M., O'Leary, A., Harro, J., Reif, A., & Freudenberg, F. (2016). Interaction of NOS1AP with the NOS-I PDZ domain: Implications for schizophrenia-related alterations in dendritic morphology. European neuropsychopharmacology: the journal of the European College of Neuropsychopharmacology, 26(4), 741–755.

5. Carrel, D., Du, Y., Komlos, D., Hadzimichalis, N. M., Kwon, M., Wang, B., Brzustowicz, L. M., & Firestein, B. L. (2009). NOS1AP regulates dendrite patterning of hippocampal neurons through a carboxypeptidase E-mediated pathway. The Journal of neuroscience: the official journal of the Society for Neuroscience, 29(25), 8248–8258.

6. Cheng, C., Li, X., Gao, S., Niu, S., Chen, M., Qin, J., Guo, Z., Zhao, J., & Shen, A. (2008). Expression of CAPON after spinal cord injury in rats. Journal of molecular neuroscience: MN, 34(2), 109–119.

7. Clattenburg, L., Wigerius, M., Qi, J., Rainey, J. K., Rourke, J. L., Muruganandan, S., Sinal, C. J., & Fawcett, J. P. (2015). NOS1AP Functionally Associates with YAP To Regulate Hippo Signaling. Molecular and cellular biology, 35(13), 2265–2277.

8. Coon, E. A., & Singer, W. (2020). Synucleinopathies. Continuum (Minneapolis, Minn.), 26(1), 72.

9. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... & Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience, 10(2), giab008.

10. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15-21.

11. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047-3048.

12. Freudenberg, F., Alttoa, A., & Reif, A. (2015). Neuronal nitric oxide synthase (NOS1) and its adaptor, NOS1AP, as a genetic risk factors for psychiatric disorders. Genes, brain, and behavior, 14(1), 46–63.

13. Hashimoto, S., Matsuba, Y., Kamano, N., Mihira, N., Sahara, N., Takano, J., Muramatsu, S. I., Saido, T. C., & Saito, T. (2019). Tau binding protein CAPON induces tau aggregation and neurodegeneration. Nature communications, 10(1), 2394.

14. Lawford, B. R., Morris, C. P., Swagell, C. D., Hughes, I. P., Young, R. M., & Voisey, J. (2013). NOS1AP is associated with increased severity of PTSD and depression in untreated combat veterans. Journal of affective disorders, 147(1-3), 87-93.

15. Li, L. L., Ginet, V., Liu, X., Vergun, O., Tuittila, M., Mathieu, M., Bonny, C., Puyal, J., Truttmann, A. C., & Courtney, M. J. (2013). The nNOS-p38MAPK pathway is mediated by NOS1AP during neuronal death. The Journal of neuroscience: the official journal of the Society for Neuroscience, 33(19), 8185–8201.

16. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), 1-21.

17. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., ... & Cunningham, F. (2016). The ensembl variant effect predictor. Genome biology, 17(1), 1-14.

18. Miranda, A., García, J., López, C., Gordon, D., Palacio, C., Restrepo, G., Ortiz, J., Montoya, G., Cardeño, C., Calle, J., López, M., Campo, O., Bedoya, G., Ruiz-Linares, A., & Ospina-Duque, J. (2006). Putative association of the carboxy-terminal PDZ ligand of neuronal nitric oxide synthase gene (CAPON) with schizophrenia in a Colombian population. Schizophrenia research, 82(2-3), 283–285.

19. Wang, J., Jin, L., Zhu, Y., Zhou, X., Yu, R., & Gao, S. (2016). Research progress in NOS1AP in neurological and psychiatric diseases. Brain research bulletin, 125, 99–105.

20. Zheng, Y., Li, H., Qin, W., Chen, W., Duan, Y., Xiao, Y., Li, C., Zhang, J., Li, X., Feng, G., & He, L. (2005). Association of the carboxyl-terminal PDZ ligand of neuronal nitric oxide synthase gene with schizophrenia in the Chinese Han population. Biochemical and biophysical research communications, 328(4), 809–815.

21. Zhu, A., Ibrahim, J. G., & Love, M. I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics, 35(12), 2084-2092.

УДК 577.22

Potential cancer dependencies in the context of LKB1 loss in non-small cell lung cancer

T. Kikalova, A. Holik

Discovery & Translational Science, Clarivate, Calle Provenza 398, Barcelona 08025, Spain.

Non-small-cell lung carcinoma (NSCLC) is any type of epithelial lung cancer other than small-cell lung carcinoma. NSCLC accounts for about 85% of all lung cancers and is known to be relatively insensitive to chemotherapy, comparing to small-cell carcinoma [1, 2].

NSCLC employs multiple ways to avoid apoptosis and gain chemoresistance, that makes the treatment more challenging [3]. But despite broad interest and active research in this area, the progress with identifying the effective treatment is still limited.

One of the molecular characteristics of NSCLC is a frequent loss of the tumor-suppressor kinase LKB1 (liver kinase B1). LKB1 is known for its ability to induce apoptosis, regulate cell polarity and differentiation and suppress the growth, invasion, and metastases of tumor cells [4, 5].

Although the inhibition of tumor-suppressors, such as LKB1, gives an advantage in avoiding the apoptosis, it also affects the normal pathways and thus the tumor cells may have to rely on alternative means of survival. This gives us an opportunity to identify effective targets in these alternative pathways that we can inhibit and by this affect only tumor cells without damaging normal tissues [6], an approach commonly referred to as synthetic lethality.

The main goal of this project was to identify genes in alternative pathways of cancer survival in condition of LKB1 loss and analyze identified genes for their safety and success as potential drug target candidates.

To discover these alternative pathways, open data from DepMap (the Broad Institute cancer Dependency Map project, version 22Q1) and TCGA (Cancer Genome Atlas Project) databases with defective and functional LKB1 was compared.

In Phase 1 the data on sensitivity to genetic targeting (survival of cell lines under condition of knockout by shRNA or CRISPR), mutations and gene expression in cancer cell lines from the DepMap project was analyzed. Lung cancer cell lines were divided into 2 groups: NSCLC cell lines with loss of LKB1 (LKB1-) and all other types of lung cancer cell lines with functional LKB1 (LKB1+). NSCLC cell lines insensitive to LKB1 knockout and with either reduced expression or damaging mutations in *LKB1*, were determined as LKB1- group. In the obtained two groups of cell lines, expression and sensitivity to knockout was compared for each gene,

excluding cell lines containing damaging mutations in these genes. As a result of Phase 1 analysis, 4 candidate genes were identified: *ONECUT3* (EntrezGene ID: 390874), *AHR* (EntrezGene ID: 196), *ERF* (EntrezGene ID: 2077), and *NR2F6* (EntrezGene ID: 2063).

In Phase 2 of the project, the clinical data from the TCGA database was explored. The cases from the TCGA-LUAD project (lung adenocarcinoma) with "lung" primary site were taken into analysis.

The clinical data was divided by mutation status. Cases containing damaging mutations in the LKB1 gene were identified as the LKB1- cohort, while all other cases - as the LKB1+ cohort. Gene expression was compared between cohorts using the RNA-seq data. As a result of the analysis and comparison of clinical data, one candidate gene (*NR2F6*) obtained in the Phase 1 of the project, was confirmed.

The goal of Phase 3 was to deeply analyze the results of the first two phases of the project using literature search and scientific databases.

The results of literature analysis supported the role of *NR2F6* as a novel therapeutic target for lung adenocarcinoma treatment. The inhibition of *NR2F6* was shown to suppress proliferation, migration and invasion of lung adenocarcinoma [7, 8].

Despite the fact that other candidate genes from Phase 1 were not confirmed by clinical data, there are multiple studies showing a significant role of the *AHR* [9, 10] and *ERF* [11, 12] genes in the development of lung adenocarcinoma. And thus, these genes might also constitute promising drug target candidates in the context of LKB1 loss.

To identify the molecular pathways affected by the loss of LKB1 in NSCLC, the following analysis was performed:

- detection of the overexpressed genes associated with the loss of LKB1 (548 genes were detected);

- identification of the common group of downstream genes indirectly regulated by both LKB1 and the confirmed candidate gene, *NR2F6,* using MetaCore™ network reconstruction toolkit (61 genes were identified);

- the GO process enrichment analysis was made for both gene sets (overexpressed genes and common downstream genes) and the resultant GO processes were overlapped to obtain the intersection.

As a result of the intersection, 13 common processes were identified that might be grouped into 3 logical high-level groups of processes associated with: cellular response, regulation of cellular metabolism and regulation of apoptosis:

- response to organic substance
- cellular response to chemical stimulus

- response to oxygen-containing compound
- cellular response to organic substance
- response to endogenous stimulus
- response to lipid
- positive regulation of biological process
- cellular nitrogen compound biosynthetic process
- positive regulation of cellular process
- positive regulation of metabolic process
- negative regulation of biological process
- regulation of cell death
- regulation of apoptotic process

Thus, the processes that are affected by the loss of LKB1 and that are compensated by the increase in the expression of substitute genes were identified.

In conclusion, the successfully confirmed gene-candidate NR2F6 is recommended for further research as a drug target for the NSCLC. Two other genes identified in Phase 1 (AHR and ERF) are also interesting candidates for future research. The identified commonly regulated processes should be taken into consideration during the future drug toxicity research and compensatory therapy development.

Details on the results and workflow can be found in the GitHub repository: https://github.com/Tatiana-kik/NSCLC_dependencies_LKB1.

References

1. Ettinger DS. Overview and state of the art in the management of lung cancer. Oncology (Williston Park). 2004 Jun;18(7 Suppl 4):3-9. [PMID: 15255162]

2. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc. 2008 May;83(5):584-94. doi: 10.4065/83.5.584. [PMID: 18452692]

3. Chang A. Chemotherapy, chemoresistance and the changing treatment landscape for NSCLC. Lung Cancer. 2011 Jan;71(1):3-10. doi: 10.1016/j.lungcan.2010.08.022. Epub 2010 Oct 16. [PMID: 20951465]

4. Mograbi B, Heeke S, Hofman P. The Importance of STK11/LKB1 Assessment in Non-Small Cell Lung Carcinomas. Diagnostics (Basel). 2021 Jan 29;11(2):196. doi: 10.3390/diagnostics11020196. [PMID: 33572782]

5. Mazzaschi G, Leonetti A, Minari R, Gnetti L, Quaini F, Tiseo M, Facchinetti F. Modulating Tumor Microenvironment: A Review on STK11 Immune Properties and Predictive vs Prognostic Role for Non-small-cell Lung Cancer Immunotherapy. Curr Treat Options Oncol. 2021 Sep 15;22(11):96. doi: 10.1007/s11864-021-00891-8. [PMID: 34524570]

6. Nawaz K, Webster RM. The non-small-cell lung cancer drug market. Nat Rev Drug Discov. 2016 Apr;15(4):229-30. doi: 10.1038/nrd.2016.42. [PMID: 27032828]

7. Jin C, Xiao L, Zhou Z, Zhu Y, Tian G, Ren S. MiR-142-3p suppresses the proliferation, migration and invasion through inhibition of NR2F6 in lung adenocarcinoma. Hum Cell. 2019 Oct;32(4):437-446. doi: 10.1007/s13577-019-00258-0. Epub 2019 Jun 5. [PMID: 31168689]

8. Klepsch V, Siegmund K, Baier G. Emerging Next-Generation Target for Cancer Immunotherapy Research: The Orphan Nuclear Receptor NR2F6. Cancers (Basel). 2021 May 26;13(11):2600. doi: 10.3390/cancers13112600. [PMID: 34073258]

9. Chang JT, Chang H, Chen PH, Lin SL, Lin P. Requirement of aryl hydrocarbon receptor overexpression for CYP1B1 up-regulation and cell growth in human lung adenocarcinomas. Clin Cancer Res. 2007 Jan 1;13(1):38-45. doi: 10.1158/1078-0432.CCR-06-1166. [PMID: 17200336]

10. Cheng YH, Huang SC, Lin CJ, Cheng LC, Li LA. Aryl hydrocarbon receptor protects lung adenocarcinoma cells against cigarette sidestream smoke particulates-induced oxidative stress. Toxicol Appl Pharmacol. 2012 Mar 15;259(3):293-301. doi: 10.1016/j.taap.2012.01.005. Epub 2012 Jan 16. [PMID: 22273977]

11. Chou YT, Lin HH, Lien YC, Wang YH, Hong CF, Kao YR, Lin SC, Chang YC, Lin SY, Chen SJ, Chen HC, Yeh SD, Wu CW. EGFR promotes lung tumorigenesis by activating miR-7 through a Ras/ERK/Myc pathway that targets the Ets2 transcriptional repressor ERF. Cancer Res. 2010 Nov 1;70(21):8822-31. doi: 10.1158/0008-5472.CAN-10-0638. Epub 2010 Oct 26. [PMID: 20978205]

12. Liu Z, Zheng M, Lei B, Zhou Z, Huang Y, Li W, Chen Q, Li P, Deng Y. Whole-exome sequencing identifies somatic mutations associated with lung cancer metastasis to the brain. Ann Transl Med. 2021 Apr;9(8):694. doi: 10.21037/atm-21-1555. [PMID: 33987392]

УДК 004.94, 577.22

Correlation between DNA sequence and chromatin structure

K. Kirilenko[1], I. Kozlov[2], G. Zakharov[3]

[1]*Tomsk State University, Lenin Ave, 36, 634050, Tomsk, Russia*
[2] *ITMO University, Kronverkskiy Prospekt, 49, 197101, St Petersburg, Russia*
[3]*EPAM Systems*

Modern methods for *in silico* predicting the structure of 3D genome organization are mostly based on proteins associated with DNA. These proteins are crucial factors in chromatin formation and therefore in 3D genome organization, but there are a lot of information kept in primary DNA structure that can be relevant to chromatin formation. In this work we tried to answer the question whether DNA sequence itself can be a predictor of 3D nuclear structure.

Today, Hi-C is considered a state-of-the-art method for analyzing 3D genome organization. We chose to predict the Hi-C matrix (degree of interaction between two sites) as a reflection of the 3D genome.

DNA sequence contains a lot of information such as repetitive DNA, genes, GC-content, distance between sites, etc. It is known that euchromatin is localized in the nuclear interior and heterochromatin at the nuclear periphery, and heterochromatin is made up of repetitive DNA mainly [1]. Formation of topologically associating domains (TADs) depends on distance between different sites of DNA (if sites are close there are most likely in the same TAD) and sequence in DNA, genes with similar function which activity depends on the same enhancer more often are in the same TAD [2]. There is evidence that identical DNA sequences (30-60 nucleotides in length) play a key role in ectopic pairing of different chromosomes of *Drosophila melanogaster* [3]. These assumptions allowed us to create a program called NAP – 3D genome organization predictor.

We worked with the *Anopheles merus* genome and Hi-C data. First, we used RepeatModeler and RepeatMasker to annotate the genome for different types of repetitive DNA [4]. We used Augustus to annotate the genome for genes and other structural elements [5]. The Homology Segment Analysis program was chosen to create a similarity matrix [3]. Having all these data for the genome of *Anopheles merus* we created a ML-model.

This model predicted the interaction between two DNA sites in the nucleus, based only on data that can be obtained from the primary DNA sequence. We have built a table, each row of which contains information about different types of repetitive DNA, gene annotation and other structural elements annotation in two bins (sites), as well as information about the degree of homology between them. Gradient boosting was chosen as the machine learning method, namely the catboost library [6]. We trained and tested all ml-models at the beginning on chromosome 2R of

Anopheles merus, which contains many repeats, and then tested it on chromosome 3L of *Anopheles merus*. We have built a binary classification model that predicts either the presence or absence of interaction between 2 bins, while obtaining an f1-score equal to 0.88. We also divided the range of interaction values into 4 classes: no interaction, low interaction, medium interaction, high interaction. After that, we built a multiclass classification model that covers one of these 4 classes.

Among the results of the model analysis, it can be noted that with an increase in the distance between the bins, the degree of interaction decreases. We also noticed that the length of the coding region and the degree of homology in bins positively correlates with the degree of interaction. On this model, for these 4 classes of interactions, we got f1-score: not: 0.96, little: 0.63, middle:0.79, high: 0.84. The regression model ignores all features except the distance between the bins, which is conceptually wrong - which we are going to fix in the future.

Summing up, we can say that the primary DNA sequence is certainly a qualitative predictor of the 3D organization of chromatin.

Repetitive elements, encoding regions and the degree of homology play an important role in chromatin interactions.

References:

1. Falk, Martin, et al. "Heterochromatin drives compartmentalization of inverted and conventional nuclei." Nature 570.7761 (2019): 395-399.

2. Dixon, Jesse R., David U. Gorkin, and Bing Ren. "Chromatin domains: the unit of chromosome organization." Molecular cell 62.5 (2016): 668-680.

3. Zhuravlev, Aleksandr V., et al. "Chromatin Structure and “DNA Sequence View”: The Role of Satellite DNA in Ectopic Pairing of the Drosophila X Polytene Chromosome." International Journal of Molecular Sciences 22.16 (2021): 8713.

4. Chen, Nansheng. "Using Repeat Masker to identify repetitive elements in genomic sequences." Current protocols in bioinformatics 5.1 (2004): 4-10.

5. Mario Stanke, Ana Tzvetkova, Burkhard Morgenstern (2006) AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome BMC Genome Biology, 7(Suppl 1): S11.

6. Anna Veronika Dorogush, Andrey Gulin, Gleb Gusev, Nikita Kazeev, Liudmila Ostroumova Prokhorenkova, Aleksandr Vorobev "Fighting biases with dynamic boosting". arXiv:1706.09516, 2017

УДК 004.94, 57.08

In silico modeling of coverage profiles for multiplex target panels

A. Kislova[1], I. Pyankov[2]

[1] *Saint Petersburg State University, Universitetskaya emb 7-9, 199034, Saint Petersburg, Russia*
[2] *Parseq Lab, Russian Federation, St. Petersburg, doroga v Kamenku 74*

The development of multiplex target panels for polymerase chain reaction means that highly specific primers are designed to minimize the number of amplicons for target regions. The panels are always validated *in vitro*, but *in silico* validation would improve the existing pipeline.

The goal of this project was to test the existing tool called DegenPrimer, which was initially developed to use on small prokaryotic genomes and small amount of primers, and try to adjust it for *in silico* validation of designed target panels and check the output correlation with the real data.

DegenPrimer, developed in 2015 by Evgeniy Taranov (https://github.com/allista), performs sophisticated analysis of degenerate primers, including calculation of melting temperatures, prediction of stable secondary structures and primer dimers, cycle-by-cycle PCR simulation with any number or primers and matrices, primer specificity checks with automated BLAST queries and consequent PCR simulation using BLAST results as matrices, simulation of electrophoresis and automated optimization of PCR conditions.

We successively launched the tool on two pools of primers and an amplicons sequence, gene (CFTR) and chromosome (7th chromosome) with additional pseudogene sequence from the 20th chromosome. To primarily check the accuracy of the tool alignment mechanism and search engine, we launched the tool also using BLAST queries to the NCBI database. Then we compared the results of the DegenPrimer predictions with our real lab data and checked correlation between the concentrations of the products and the amplicons coverage profiles.

Our results show that the predicted primers, duplexes and PCR products do not fully match the real data - the accuracy of the predictions varied from 60 to 75%. When using real primers concentrations for the analysis, the tool predicts quick and full saturation of the system, which is not confirmed by the laboratory data. We did not find any correlation between the predicted product concentrations and amplicons coverage profiles. Consequently, it was decided that this tool is not suitable for *in silico* validation of the multiplex target panels.

УДК 004.94, 57.08

Generation of possible single-nucleotide variants with a given effect on protein-coding sequence

O. Kolpakova[1], Y. Barbitov[1], M. Skoblov[2]

[1] *Bioinformatics Institute, St. Petersburg, Russia*
[2] *Research Centre for Medical Genetics, Moscow, Russia*

Motivation

There are several public databases that collect data on clinically relevant genetic variants that cause phenotype changes and monogenic disease. Of these, ClinVar is the most widely used database. Variants in the coding regions of the human genome, especially missense substitutions, are the most common cause of genetic pathology. However, not all clinically significant variants have been identified and described. This complicates the identification of the molecular cause of a genetic disorder in people with a suspected hereditary disease. Given this limitation, we aimed to create a tool for generation possible new pathogenic variants in well-known disease genes (from the OMIM database) by creating a comprehensive list of all single-nucleotide variants that results in the same amino acid substitution as known pathogenic variants. This tool expands the list of possible pathogenic variants and can be used to improve the molecular genetic diagnosis of hereditary diseases.

Materials and methods

For this project we used publicly available software (Ensembl Variant Effect Predictor (VEP) [1], IGV 2.11.9 [4]) and databases (ClinVar [2], OMIM [3], gnomAD 2.1.1 [5]) and own Python script. The script is available at https://github.com/OxanaKolpakova/new_SNP.

The VCF file provided by ClinVar (GRCh38-based) has been VEP-annotated with the following flags: --cache --refseq --canonical (12858671 annotations were obtained). Using a Python script we extracted variants in OMIM genes that were missense variants in canonical isoforms. The resulting dataset contained the following numbers of variants by pathogenicity class: pathogenic - 2420, likely pathogenic - 2750, benign - 1510, likely benign - 2367, others - 54336. A custom script was used for creation of new missense variants leading to the same amino acid substitution as known ones (63383 variants were created)

Results and Discussion

We created the tool for generation of possibly pathogenic variants that lead to the same amino acid substitutions as known ones. The script's accuracy was validated on coordinates of reference SNPs by visual inspection in IGV.

New clinically significant variants belonged to the following classes (according to the source variant's class): benign - 815, likely benign - 1609, pathogenic - 876, likely pathogenic - 994. The mean gnomAD frequency of generated possibly pathogenic and likely pathogenic variants was significantly lower than of possible new benign and likely benign variants: 0.035, 0.05 and 0.005, 0.0001, respectively. This indicates that selection is directed against these variants, validating their possible role in hereditary diseases

New generated SNPs could increase the accuracy of molecular genetic diagnosis of diseases associated with OMIM genes. This tool can be used to create missense SNPs for other gene lists and variant databases, such as HGMD.

In the future, we plan to expand the capabilities of the tool by improving handling of the exon/intron boundaries, DNA strand, and splicing. We are also planning to apply the script to real medical data to identify new pathogenic SNPs.

References

1. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biology. Jun 6. 17(1):122. 2016. doi:10.1186/s13059-016-0974-4

2. McKusick's online Mendelian inheritance in man (OMIM(R)) Nucleic Acids Research. Nov 37. D793-6. 2008. doi:10.1093/nar/gkn665

3. Landrum, MJ, Lee, JM, Benson M, Brown, G. Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Research. 44. D1: D862–D868. 2016. doi:10.1093/nar/gkv1222

4. Robinson J, Thorvaldsdóttir H, Winckler W et al. Integrative genomics viewer. Nat Biotechnol. 29. 24–26. 2011. doi :10.1038/nbt.1754

5. Koch, Linda. Exploring human genomic diversity with gnomAD. Nature Reviews Genetics. 21.8: 448-448. 2020

УДК 004.94, 57.08

Analysis of variable evolutionary constraint within a single ORF

O. Kotovskaya[1], Y. Barbitoff[2], M. Skoblov[3]

[1] *Peter the Great St.Petersburg Polytechnic University, Polytechnicheskaya, 29, 195251, Saint-Petersburg, Russia*
[2] *Bioinformatics Institute, Kantemirovskaya street, 2A, 197342, Saint-Petersburg, Russia*
[3] *Research Center of Medical Genetics, Moskvorechye St, 1, 115522, Moscow, Russia*

email: kotovskaya.oa@outlook.com

Genetic protein truncation variants (PTVs) often lead to diseases if the protein is functionally important. Catalogs of human exome and genome variation have been recently constructed, which is of great importance in clinical diagnostic: for example, these resources could be used to find clinically significant genes as such genes are typically enriched with *de novo* mutations and have a low frequency of PTVs in the population (Cassa et al., 2017). At the same time, confident classification of PTVs as pathogenic is compromised by the fact that PTV variants are also found in large numbers in healthy people's genomes. Approaches based on the search for genes that do not tolerate the presence of PTV allow to identify potentially harmful mutations: the absence of functional variants may indicate the presence of evolutionary constraint leading to the removal of such variants from the population by purifying selection.

In this work, we focused on genes that could not be accurately classified as conserved (that is, under selection) or non-conserved (that is, free from selection). These mostly comprise cases when non-conserved regions are found in relatively conserved genes. This work is devoted to implementation of algorithm to the search for such sequences. We built a hidden Markov model (HMM), which allows to determine the degree of conservation of individual regions of the protein-coding sequence (CDS).

HMM is a statistical model in which a system is represented as a Markov process with hidden states generating observable states. For the sake of simplicity, we have built a model in which only two states are possible: conserved (*Cons*) and non-conserved (*Not*) (in the future, the number of states can be increased). We divide the CDS of a gene into regions of fixed length (windows). We did not fix window size for all genes, because the length of genes varies greatly. .

We choose the PTVs allele count per window obtained from the gnomAD database v.2.1.1 (Karczewski et al., 2020), n, as observations in this model (since it is finite, we can consider this quantity discrete). To assess the constraint, we used an estimation of the selective effect against heterozygous PTVs (s_{het}) that considers for each region the observed count of protein truncation variants (PTV) n, the allele

number or sample size (N), and the expected mutation rate. The mutation rate U was calculated considering the trinucleotide context for each codon. The choice is motivated by earlier findings published by Samocha et al., 2014, who showed that the best context for determining the variability of a single nucleotide is the inclusion of both 5' and 3' flanking nucleotides. We used mutation frequencies for each possible variant (G, T, C for A), specified in Supplement Materials to the work of Karczewski et al., 2020. For each nucleotide in the window, we considered all three possible substitutions. If the corresponding substitution leads to a PTV, its expected rate in a given trinucleotide context is summed up to yield per-window value of U.

Similarly, to the work of Cassa et al., 2017, we assumed the observed distribution of PTV counts across i-th region:

$$P(n_i \mid \alpha,\ \beta; \nu_i) = \int P(n_i | s_{het}; \nu_i) P\left(s_{het}; \alpha,\ \beta\right) ds_{het},$$

where $\nu_i = N_i U_i$ is expected per=gene PTV counts.

Further, because PTV mutations are rare events for genes under negative selection, observed numbers of PTV obey a Poisson distribution: $P(n_i | s_{het}; \nu_i) = Pois(n_i, \lambda_i)$, where mean $\lambda_i = \frac{\nu_i}{s_{het}}$. $P(s_{het}; \alpha,\ \beta)$ is parametrized by inversive Gaussian distribution $IG(s_{het}; \alpha,\ \beta)$ with mean α and shape β parameters which calculated for the gene as a whole. using gnomAD data (values were taken from Skitchenko et al., 2020).

Thus,

$$P(n_i \mid \alpha,\ \beta; \nu_i) = \int Pois\left(n_i,\ \frac{\nu_i}{s_{het}}\right) IG\left(s_{het}; \alpha,\ \beta\right) ds_{het}$$

Emission probabilities ($e_k(n_i)$ for the observed state n_i were obtained by the following integration for each hidden state k:

$$e_k\left(n_i\right) = \frac{P(n_i \mid \alpha, \beta; \nu_i; a, b)}{P(n_i \mid \alpha, \beta; \nu_i)} = \frac{\int_a^b Pois\left(n_i, \frac{\nu_i}{s_{het}}\right) IG\left(s_{het}; \alpha, \beta\right) ds_{het}}{P(n_i \mid \alpha, \beta; \nu_i)},$$

where $a,\ b = [0,\ 0.01]$ for $k = Not$ and $a,\ b = [0.01,\ 1]$ for $k = Cons$. The choice of such values is also due to the results obtained in the work Cassa, et al., 2017.

Since we had no assumptions about the transition probabilities, we used the Baum–Welch algorithm to find transition probabilities corresponding to the

maximum likelihood of the model. The decoding of the path of states was carried out using the Viterbi algorithm, it consists in finding a path that also meets the maximum likelihood of the model.

As a result, we implemented a complete pipeline for finding conserved regions for one gene using Snakemake v6.10.0 (Köster et al., 2021).The algorithm was tested on a set of genes: mostly conserved, mostly non-conserved, presumably possessing non conserved regions. Household genes, for which there is a high degree of constraint were selected as conserved genes (*TOP2A, HSPB8*). The genes listed as genes with protein changing signature in the PopHumanScan catalog (Murga-Moreno et al., 2019) were selected as genes with high variability (like *ZNF69*). The genes that were found earlier in the work of Skitchenko et al., 2020 were selected as genes with non-conserved regions (*ARFGEF1, PAX3, GDNF*).

Algorithm successfully solves the first two cases and turns out to be less sensitive for the third: in the case of many alleles, the algorithm with the specified parameters copes with finding the most non-conserved region. In the future, we plan to find a more rigorous approach to transition probabilities in the model, and improve handling of regions where no PTV is observed.

References

1. Cassa, C. A., Weghorn, D., Balick, D. J., Jordan, D. M., Nusinow, D., Samocha, K. E., O'Donnell-Luria, A., MacArthur, D. G., Daly, M. J., Beier, D. R., & Sunyaev, S. R. (2017). Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nature Genetics 2017 49:5, 49(5), 806–810. https://doi.org/10.1038/ng.3831

2. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … Daly, M. J. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020 581:7809, 581(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

3. Köster, J., Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., & Nahnsen, S. (2021). Sustainable data analysis with Snakemake. F1000Research, 10. https://doi.org/10.12688/F1000RESEARCH.29032.2/DOI

4. Murga-Moreno, J., Coronado-Zamora, M., Bodelón, A., Barbadilla, A., & Casillas, S. (2019). PopHumanScan: the online catalog of human genome adaptation. Nucleic Acids Research, 47(D1), D1080–D1089. https://doi.org/10.1093/NAR/GKY959

5. Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook, E. H., … Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. Nature Genetics 2014 46:9, 46(9), 944–950. https://doi.org/10.1038/ng.3050

6. Skitchenko, R. K., Kornienko, J. S., Maksiutenko, E. M., Glotov, A. S., Predeus, A. v., & Barbitoff, Y. A. (2020). Harnessing population-specific protein truncating variants to improve the annotation of loss-of-function alleles. BioRxiv, 2020.08.17.254904. https://doi.org/10.1101/2020.08.17.254904

УДК 004.94, 577.22

Construction of SARS-CoV-2 neutralizing ligands with tight binding to spike protein

A. Kovalenko[1,3], X. Sukhanova[2,3], O. Lebedenko[3], N. Skrynnikov[3]

[1] *Laboratory of Chemoinformatics, Infochemistry Scientific Center, ITMO University, Kronverkskiy Prospekt 49, 197101, Saint-Petersburg, Russia*
[2] *Laboratory of Applied genomics, SCAMT Institute, ITMO University, Kronverkskiy Prospekt 49, 197101, Saint-Petersburg, Russia*
[3] *Laboratory of Biomolecular NMR, Saint Petersburg State University, Universitetskaya emb 7-9, 199034, Saint Petersburg, Russia*

Introduction

Over the last decade, the emergence of *in silico* tools has paved the way to rational drug discovery. A number of computational protocols have been developed for *de novo* design of high-affinity binders based on target structure alone. In response to COVID19 pandemic, these methods were used to engineer mini-proteins (MPs) capable of blocking the receptor-binding domain of spike protein from the wild-type virus (RBD-wt). However, since the initial discovery a number of SARS-CoV-2 variants have emerged, including some highly transmissible strains that proved responsible for millions of deaths around the globe. In this study, we provide the computational scheme to assess the binding affinity of the most promising mini-proteins, MP1 and MP3, against the RBD of the newer variants of coronavirus, delta+ and omicron. For this purpose, we applied the well-established protocol based on molecular mechanics/generalized Born surface area (MM/GBSA) method to calculate the difference in binding energy ΔΔG between RBD of the new variants (delta+, omicron) and RBD-wt in complex with MP1 or MP3. As a step further, we also proposed the optimized version of MP3 carrying a single point mutation D37R, which shows increased affinity to RBD-delta+. This suggestion is supported by ΔΔG predictions using Flex ddG module from the modeling suite Rosetta, as well as by MM/GBSA calculation using Amber 20 platform.

Methods

The starting coordinates for complexes of MP1 and MP3 with the RBD-wt have been taken from the PDB structures 7JZU and 7JZM, respectively. To produce models of complexes of MP1 and MP3 with newer RDB variants (delta+, omicron), we performed *in silico* mutagenesis using Biobb python library (v 3.9.4). All coordinates were than regularized by geometry minimization tool from the Phenix software (v 1.19.2). The resulting models were used to start molecular dynamic (MD) trajectories in TIP4P-Ew water. All trajectories were recorded under ff14SB force field using Amber 20 MD simulation package. The simulations were conducted using NPT ensemble with Bussi thermostat. The length of each trajectory was 1.5 μs. All complexes modeled in this study remain structurally unchanged during the course of

simulations (rmsd of C^{α} atomic coordinates does not exceed ~2Å). The first 500 ns of each trajectory were interpreted as an equilibration period and discarded. The 1000 frames from the remaining part of each trajectory (500-1500 ns) were used as an input for MM/GBSA calculations. The MD data were extracted from the restart files, unwrapped and centered by means of CPPTRAJ utility of Amber. The topology files were prepared using ante-MMPBSA.py script from the Amber suite. Free energy calculations using implicit solvent GBNeck2 (igb=8) with atomic radii set mbondi3 were conducted using the script MMPBSA.py (employing the solvent dielectric constant of 80, surften parameter of 0.0072 $kcal \cdot mol^{-1} \cdot Å^{-2}$ and salt concentration 150 mM). Finally, a pairwise energy decomposition was conducted using the option idecomp=4 in MMPBSA.py with the goal to identify key binding interactions.

To optimize the sequence of MP1 and MP3 proteins for stronger binding to RBD-delta+ or RBD-omicron, we used the *in silico* saturation mutagenesis as implemented in the Flex ddG protocol of Rosetta. The protocol was used with default settings (35 structures, 35,000 backrub steps).

Structure manipulations, analyses and visualization were aided by in-house python scripts, employing the libraries pyxmolpp2 (v 1.6.0) and amber-runner (v 0.0.8), as well as in-house library md-utils, which is an extension of pyxmolpp2 with a number of additional functions for structure analyses.

Results

Our MM/GBSA calculations predict a significant drop in affinity of MP1 to RBD-delta+ as well as RBD-omicron (ΔΔG of 12.1 and 24.9 kcal/mol, respectively). The pairwise residue energy decomposition has revealed that the weakened binding of MP1 to RBD-delta+ is mainly due to mutation K417N. This mutation causes a loss of the two salt bridges, involving residues R403 and K417 on the target side and residue D30 on the ligand side. In addition to K417N mutation, the RBD-omicron carries several other mutations, of which N501Y has the most pronounced effect. The insertion of large aromatic residue in this position leads to a slight shift of MP1 relative to RBD. As a result, the geometry of all contacts changes and they generally become less favorable. The saturation mutagenesis scan using Flex ddG failed to identify any single-point MP1 mutation that could potentially improve its affinity to RBD-delta+ or RBD-omicron.

As for the MP3 mini-protein, the MM/GBSA calculations predict only a moderate decrease in binding affinity to RBD-delta+ (ΔΔG = 3.3 kcal/mol). Pairwise residue energy decomposition suggests that the effect stems from the RBD mutation K417N, which disrupts the original polar interaction with residue D11 in MP3. Of interest, in our calculations the MP3 mini-protein shows an increase in binding affinity to RBD-omicron (ΔΔG = -4.9 kcal/mol). This effect is mainly due to Q493R mutation in RBD, leading to a highly stabilizing interaction with residue D37 in MP3.

To improve the binding affinity of MP3 to RBD-delta+, we conducted an *in silico* saturation mutagenesis scan using Flex ddG module of the popular program

Rosetta. By doing so, we have identified the mutation D37R that appears to be one of the most stabilizing and does not compromise the solubility of the mini-protein. We have further investigated the binding of MP3 (D37R) to RBD-delta+ by means of the MD-based MM-GBSA analysis. These calculations predicted a substantial improvement in binding ($\Delta\Delta G$ = -4.8 kcal/mol for MP3 containing D37R mutation vs. 3.3 kcal/mol for the original-sequence MP3). Residue pairwise decomposition analysis indicated that the gain in binding affinity is due to the newly formed salt bridge, R37-E484.

This work demonstrates that the integrative *in silico* approaches can be used for fast assessment and optimization of protein therapeutic leads. In particular, the consistency between the predictions of Flex ddG and MM/GBSA methods give us a high degree of confidence that these predictions are accurate. We currently plan to test this hypothesis experimentally by measuring the binding affinity of MP3 (D37R) to RBD-delta+.

This study has been supported by grant 72777155 from St. Petersburg State University.

УДК 004.94

Analysis of the effects of combinations of single nucleotide polymorphisms within a single codon

E. Kravchuk[1], M. Skoblov[2], Y. Barbitoff[3]

[1] Lomonosov Moscow State University, 119991, 1 Leninskiye gori, Moscow, Russia
[2] Research Center of Medical Genetics, 115522, 1 Moskvorechye St, Moscow, Russia
[3] Bioinformatics Institute, Kantemirovskaya street, 2A, 197342, Saint-Petersburg, Russia

Multi-nucleotide variants (MNVs) are defined as clusters of two or more nearby variants existing on the same haplotype in an individual. When variants in an MNV are found within the same codon, the overall impact may differ from the functional consequences of the individual variants. Modern publicly available tools incorrectly annotate polymorphisms in the same codon, considering their contributions independently. It would be useful to create a tool to properly annotate MNVs.

The aim of the study was to create a tool that correctly predicts the effects of polymorphisms within a single codon.

We implemented MNVFinder, a command-line tool for searching MNVs and annotating them. It reads a VEP annotation result in a VCF format and creates a pandas.DataFrame with the data. Then it searches for SNPs within a single codon and annotates them with their combined effect on the coding sequence. The output is a tab file with the annotation if the MNVs. One can also save the tab files with annotated SNPs which are not found in the same codon with other SNPs, as well as wrongly annotated ones. We performed validation on gnomAD data (selected SNPs with population frequency 5% or more) and ClinVar data (2022-05-07 18:17).

The code and all our results can be found in the GitHub repository: https://github.com/Kravchuk-Ekaterina/MNVFinder

УДК 579.25, 577.22

Studying *Salmonella* gene expression dynamics in response to novobiocin

S. Kupriyanov[1, 2], V. I. Ladyhina[1, 4], A. A. Tkachenko[1, 3]

[1] *Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia*
[2] *Federal State Budgetary Educational Institution of Higher Education "South-Ural State Medical University" of the Ministry of Healthcare of the Russian Federation, Vorovskogo st. 64, 454092, Chelyabinsk, Russia*
[3] *ITMO University, Kronverksky Pr. 49, bldg. A, 197101, St. Petersburg, Russia*
[4] *Swedish University of Agricultural Sciences, Ulls väg 26, 75651, Uppsala, Sweden*

Introduction

Salmonella enterica is one of the most common types of enteropathogenic bacteria on Earth with more than 2500 serovars. It is known to be the cause of nontyphoidal foodborne infections (one of four key global causes of diarrheal diseases) and enteric fever in humans as well as salmonellosis in animals [1, 2]. In enteric bacteria, DNA supercoiling is very sensitive to various environmental conditions and is a sensor of various stress factors. The antibiotic novobiocin is an inhibitor of the ATPase domain of DNA gyrase; its action affects the global topology of cell DNA. The aim of this work was to study the effect of novobiocin on gene expression in *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. 14028S at various time points. To study the dynamics of changes in gene expression under the influence of novobiocin, weighted gene coexpression network analysis was used in this work. Gene co-expression networks are increasingly used to explore the system-level functionality of genes. The network construction is conceptually straightforward: nodes represent genes and nodes are connected if the corresponding genes are significantly co-expressed across appropriately chosen tissue samples [3]. In this work, we studied the effect of an antibiotic that changes the degree of DNA helix, which triggers various intracellular processes [4], so the construction of gene coexpression networks makes biological sense.

Methods and materials

Data

In this work we analyzed publicly available RNA-seq data of *Salmonella enterica* bacterial cultures treated with 500 μg of novobiocin, as well as control cultures[5]. Samples were taken at several time points: 0, 10, 20, 60 minutes for control samples; 10, 20, 60 minutes for samples treated with 500 μg of novobiocin; 60 minutes for samples treated with 100 μg of novobiocin. For each time point, 3 biological replicates were made with a total of 24 samples. Analyzed data was processed as described in the article of *Gogoleva et. al, 2020* [5] and checked for read

quality using FASTQC version 0.11.9. During the analysis we did not take into account control samples at 0 time point to ensure that the time points were equal between the novobiocin treated group and controls. Thus, for further processing, a total of 18 samples were examined. Genome sequence of the *Salmonella enterica* subsp. *enterica* serovar Typhimurium strain 14028S assembly GCA_000022165.1 was used as a reference (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/022/165/GCF_000022165.1_ASM2216v1/GCF_000022165.1_ASM2216v1_genomic.fna.gz). The genome sequence and annotation file are available at (https://www.ncbi.nlm.nih.gov/nuccore/NC_016856.1).

Methods

HISAT2 version 2.2.0 was used to build index of the reference genome and align clean reads to the genome with the following parameters: hisat2 -p --dta -x -U –S. SAM files of alignments created by HISAT2 were converted to BAM files using SAM-tools sort. Mapped reads were summarized at the transcript level into a count matrix using the assembler of RNA-Seq alignments StringTie version 2.2.1. Converting the alignment results to a matrix of read counts mapped to particular genomic features was done using the python script provided in the StringTie tutorial (http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual).

DESeq2 version 1.34.0 was used to estimate the variance between groups of samples and repeated samples using principal component analysis (PCA), as well as to normalize the number of reads. DESeq2 analysis was performed for Transcript_counts_matrix. The DESeq object was created using the formula: ~ Treated+Time+Treated:Time. Genes with log2FoldChange < 1.5 and FDR < 0.05 were considered significant.

The WGCNA version 1.70-3 package has been used. Modules of co-expressed genes were built; topological matrix type was chosen "signed"; the power parameter was calculated using the pickSoftThreshold() function as the smallest one, at which the resulting network had a similarity index with the scale-free network of more than 0.8. The relationship of modules with the time of cultivation and treatment with novobiocin was determined; Spearman's correlation coefficient was used for calculation. The key genes (hub genes) for the modules associated with antibiotic treatment were determined by relationships with the module's own genes (the first major component for the genes included in the module). Genes were recognized as key for the module if the correlation with the module's own genes was more than 0.9.

ClusterProfiler version 4.2.2 was used to perform the KEGG enrichment analysis. The significance threshold for pathways was chosen as 0.05. The modules associated with "Treated" and "Time" were mapped to the chromosome using the WoPPER online tool [6].

To perform reproducible analysis we wrapped the existing code into snakemake pipeline [7]. The full code and results can be found on Github (https://github.com/ValeriiaLadyhina/BI_Project_analysis_of_effect_of-novobiocin_on_Salmonella.git).

Results

Biological replicas cluster best with each other; according to the first component, the samples are best separated in time; according to the second component, the samples are separated by the presence of novobiocin. Genes for which log2FoldChange < 1.5 and FDR < 0.05 were considered insignificant; a total of 789 significantly differentially expressed genes were identified. out of which there are 370 upregulated genes and 419 downregulated genes. The processed samples at the time points of 10 and 20 minutes and the control samples at the same time points were grouped into two separate groups while both the control and experimental samples at time point 60 minutes were allocated into a separate group. The data demonstrate an increase in the similarity of gene expression profiles of treated and untreated samples at 60 minutes.

After building the network of gene co-expression, 10 connected modules and 1 zero module were defined. Number of genes in modules were 154, 56, 937, 927, 462, 425, 294, 179, 176, 72, and 47. Modules associated with the "Treated" variable: brown, turquoise, black, yellow. Modules associated with the "Time" variable: green, blue, red. Modules black, brown, yellow with a strong positive correlation with the "Treated" variable were combined into one group by hierarchical clustering.

Several pathways were found in both the DESeq2 assay and the modules associated with novobiocin treatment:

- C5-Branched dibasic acid metabolism;
- 2-Oxocarboxylic acid metabolism;
- Two-component system;
- Ascorbate and aldarate metabolism;
- Glycolysis / Gluconeogenesis.

One of the key concepts of WGCNA is the concept of hub genes; hub genes are genes that have a large number of connections with other genes in the module [3]; such genes are most similar to the module's eigengene (first principal component) and they have the greatest relationship with the variable under study; hub genes were considered to have a correlation coefficient with the eigengene of more than 0.9. We obtained hub genes for modules associated with novobiocin and time. Number of hub genes for different modules were: black - 42; brown - 167; yellow - 108; turquoise - 284; red - 94; blue - 147; green - 160. Biological pathways identified by both DESeq2 and hub genes analysis:

- C5-Branched dibasic acid metabolism;

- 2-Oxocarboxylic acid metabolism;
- Two-component system;
- Glycolysis / Gluconeogenesis.

These pathways are the most sensitive to novobiocin treatment; they were used to visualize the dynamics of their expression.

The topological state of DNA influences its affinity for some DNA binding proteins, especially in DNA sequences that have a high A + T base content. For example, H-NS nucleoid-associated transcription-silencing protein has been described to bind to the region of A + T-rich DNA into the promoter Pleu-500 [8]. We compared the AT composition of the hub gene sequences associated with novobiocin treatment with the AT composition of the nodal genes associated with time. The AT composition was calculated as the sum of A and T divided by the length of the sequence. This value was calculated for each gene from the list of hub genes. The results for the “Treated” and “Time” modules were combined. The distributions for the values were significantly different from normal (p-value = 2.2e-16 and p-value = 1.273e-12 for the "Treated" and "Time" modules, respectively; Shapiro-Wilk test), so the nonparametric Mann-Whitney test was used. The groups were significantly different from each other (p-value = 6.823e-11), although the difference between the means was small (0.496 and 0.470 for "Time" and "Treated", respectively). The slight difference in the AT composition can be explained by the fact that it is important for supercoiling for the promoter regions of genes, so we studied it in the initial segments of the sequence using the sliding window method. A "window" of 30 nucleotides long (the approximate length of a promoter) was iteratively shifted one nucleotide from the beginning of the gene sequence. The AT composition was calculated within the limits of the window according to the above method. For each module, a sequence of values of the AT composition averaged for the hub genes of the module within the "window" is calculated.

Discussion

Salmonella enterica is Gram-negative, food-borne pathogen that causes animal and human diseases ranging from mild to severe systemic infections. It was shown that GyrA of *Salmonella enterica* influences DNA supercoiling and as the result affects expression of stress response pathways [9]. In this work we investigated fluctuations of *Salmonella enterica* gene expression over time under treatment by novobiocin - an antibiotic that can relax DNA supercoiling and by this alter the expression of supercoiling-sensitive genes. Biological pathways may have different dynamics over time. Two-component system genes increase their expression by 20 minutes with a subsequent decrease. C5-Branched dibasic acid metabolism and 2-Oxocarboxylic acid metabolism genes increase expression over time. Glycolysis/Gluconeogenesis genes increase their expression over time in control samples, however, when treated with novobiocin, they do not show any noticeable dynamics. Two-component system is a system for perceiving changes in the environment [10]; this system can also be associated with a change in supercoiling,

which also responds to a large number of stress stimuli (pH, osmotic composition, etc.) [11]. The lack of dynamics in the expression of glycolysis pathway genes may be associated with the bacteriostatic effect of novobiocin (these processes are associated with anabolic pathways); C5-Branched dibasic acid metabolism and 2-Oxocarboxylic acid metabolism may be involved in some of the signaling pathways associated with changes in supercoiling.

The co-expression modules show a diffuse distribution along the length of the chromosome (mapping by WoPPER), which may correspond to the influence of a systemic process, such as a change in supercoiling. Genes in co-expressed modules are located at significant distances from each other (more than 1Mp). This can be explained by the topology of the chromosome, but does not exclude the influence of DNA supercoiling on the activation of modules.

The composition of A and T in the genes sensitive to novobiocin treatment was significantly greater than in the genes sensitive to time dynamics (although the difference was small). The averaged AT composition in the initial regions of the genes (the first 300 nucleotides) was higher in the genes associated with novobiocin, although it showed a similar distribution pattern (graphs can be viewed in the project Gighub repository https://github.com/ValeriiaLadyhina/BI_Project_analysis_of_effect_of-novobiocin_on_Salmonella). Thus, the AT-composition of genes sensitive to novobiocin treatment significantly differs from other genes. This is consistent with the data that the chemical composition of genes affects their sensitivity to DNA superspiralization[8]. During transfer, such genes interact with other genes sensitive to superspiralization, which allows them to be included in regulatory networks. This can play a big role in changing the functioning of cells. Therefore, the identification of genes sensitive to DNA superspiralization can be a useful tool for studying the adaptability of organisms. The determination of AT-composition has some potential for the detection of such genes.

References

1. World Health Organization Salmonella (non-typhoidal), Retrieved June 10, 2019, from https://www.who.int/news-room/fact-sheets/detail/salmonella-(non-typhoidal)

2. V T Nair, Divek et al. “Antibiotic-Resistant Salmonella in the Food Supply and the Potential Role of Antibiotic Alternatives for Control.” Foods (Basel, Switzerland) vol. 7,10 167. 11 Oct. 2018, doi:10.3390/foods7100167

3. Zhang, Bin, and Steve Horvath. “A general framework for weighted gene co-expression network analysis.” Statistical applications in genetics and molecular biology vol. 4 (2005): Article17. doi:10.2202/1544-6115.1128

4. Baranello, Laura et al. “The importance of being supercoiled: how DNA mechanics regulate dynamic processes.” Biochimica et biophysica acta vol. 1819,7 (2012): 632-8. doi:10.1016/j.bbagrm.2011.12.007

5. Gogoleva, Natalia E et al. “Transcriptomic data of Salmonella enterica subsp. enterica serovar Typhimurium str. 14028S treated with novobiocin.” Data in brief vol. 29 105297. 17 Feb. 2020, doi:10.1016/j.dib.2020.105297

6. Puccio S, Grillo G, Licciulli F, Severgnini M, Liuni S, Bicciato S, De Bellis G, Ferrari F, Peano C WoPPER: Web server for Position Related data analysis of gene Expression in Prokaryotes. Nucleic Acids Res. 2017; 45. DOI: 10.1093/nar/gkx329

6. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. F1000Res 10, 33.

7. Dorman, Charles J, and Matthew J Dorman. “DNA supercoiling is a fundamental regulatory principle in the control of bacterial gene expression.” Biophysical reviews vol. 8,Suppl 1 (2016): 89-100. doi:10.1007/s12551-016-0238-2

8. Webber, Mark A et al. “Clinically relevant mutant DNA gyrase alters supercoiling, changes the transcriptome, and confers multidrug resistance.” mBio vol. 4,4 e00273-13. 23 Jul. 2013, doi:10.1128/mBio.00273-13

9. Liu, Cong et al. “Two-Component Signal Transduction Systems: A Major Strategy for Connecting Input Stimuli to Biofilm Formation.” Frontiers in microbiology vol. 9 3279. 10 Jan. 2019, doi:10.3389/fmicb.2018.03279

10. O'Byrne, C P et al. “The DNA supercoiling-sensitive expression of the Salmonella typhimurium his operon requires the his attenuator and is modulated by anaerobiosis and by osmolarity.” Molecular microbiology vol. 6,17 (1992): 2467-76. doi:10.1111/j.1365-2958.1992.tb01423.x

УДК 004.94, 573.22

Structure-based modeling of cysteine and serine disease variants of human proteome

D. Podgalo

Smolensk State Medical University

Introduction

During the early 1980s, the ability to rationally design drugs using protein structures was an unrealized goal for many structural biologists. The first projects were underway in the mid-80s, and by the early 1990s the first success stories were published. Today, even though there is still quite a bit of fine-tuning necessary to perfect the process, structure-based drug design is an integral part of most industrial drug discovery programs and is the major subject of research for many academic laboratories.

The completion of the human genome project, the start of both the proteomics and structural genomics revolutions, and developments in information technology are fueling an even greater opportunity for structure-based drug design to be part of the success story in the discovery of new drug leads. Excellent drug targets are identified at an increased pace using developments in bioinformatics. The genes for these targets can be cloned quickly, and the protein expressed and purified to homogeneity. Advances in high-throughput crystallography, such as automation at all stages, more intense synchrotron radiation, and new developments in phase determination, have shortened the timeline for determining structures. Faster computers and the availability of relatively inexpensive clusters of computers have increased the speed at which drug leads can be identified and evaluated in silico [1].

There are many disease-associated mutations that endow pharmacological target (typically a protein) with drug resistance, e.g. G12C amino acid substitution in oncogenic target KRAS [10]. People carrying such mutations may need the development of personalized drugs that take into account structural peculiarities of the mutated protein. One of the promising strategies is to develop covalent drugs that are specific for a given mutation [9].

The goal of this project is to model structures of human proteins with disease-associated amino acid substitutions. Two types of amino acid substitutions are selected: X to cysteine or X to serine (X is any amino acid residue) – these residues are often used as the attachment points for covalent drugs.

Materials and methods

We used the following software:

1. Conda [4]

2. Ensembl Variant Effect Predictor (VEP) to predict amino acid substitution based on single nucleotide polymorphism [11]

3. Rosseta ddg_monomer to model mutant proteins [2]

4. SCons to install Rosetta ddg_monomer [6]

5. UniProt Retrieve/ID mapping to map VEP-annotated proteins ID to UniProt ID [12]

We used the following databases:

1. gnomAD 2.1.1 ExAC (a database of single nucleotide polymorphism) with 9362318 variants (60705 exomes) for human genome assembly GRCh37/hg19 [7]

2. VEP human database GRCh37/hg19 for VEP-annotation

3. ClinVar database for disease detection [8]

4. AlphaFold2 database – normal sturcture-based models of human proteins [5]

Results

We annotated all 9362318 gnomAD variants using VEP and then also used VEP to filter these variants and leave only missense mutations inside coding sequence with amino acid substitution to cysteine or serine. Also, we chose only pathogenic variants that are reliably known for the disease.

After this we got 1339 cysteine and serine variants associated with disease. Then each variant we linked with the AlphaFold2 model and using Rosseta ddg_monomer modeled mutant protein.

We also carried out statistical processing of the results. Most diseases are associated with proteins: DYHC2 (8.36 %), USH2A (7.84 %), VWF (3.36 %). The most frequently substituted amino acid is arginine (52.05 %), also commonly substituted amino acids are glycine (14.71 %) and tyrosine (10.9 %). In 66.69 %, the substitution led to the appearance of cysteine, in 33.31 % - of serine. Most frequent diseases associated with amino acid substitution in proteins are asphyxiating thoracic dystrophy (6.05 %), primary ciliary dyskinesia (4.03 %) and retinal dystrophy (4.03 %).

All our results and pipeline of this work can be found in the GitHub repository: DmitriiPodgalo/POP.

Discussion

The obtained structural models will be used as the starting conformations for the structure-based drug design pipelines [3].

Structure-based drug design is a powerful method, especially when used as a tool within an armamentarium, for discovering new drug leads against important targets. After a target and a structure of that target are chosen, new leads can be designed from chemical principles or chosen from a subset of small molecules that scored well when docked in silico against the target. After a preliminary assessment of bioavailability, the candidate leads continue in an iterative process of reentering structural determination and reevaluation for optimization. Focused libraries of synthesized compounds based on the structure-based lead can create a very promising lead which can continue to phase I clinical trials.

As structural genomics, bioinformatics, and computational power continue to explode with new advances, further successes in structure-based drug design are likely to follow. Each year, new targets are being identified, structures of those targets are being determined at an amazing rate, and our capability to capture a quantitative picture of the interactions between macromolecules and ligands is accelerating [1].

References

1. Anderson, Amy C. "The process of structure-based drug design." Chemistry & biology 10.9 (2003): 787-797.

2. Barlow, Kyle A., et al. "Flex ddG: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation." The Journal of Physical Chemistry B 122.21 (2018): 5389-5399.

3. Cohen, Nissim Claude. "Structure-based drug design and the discovery of aliskiren (Tekturna): perseverance and creativity to overcome a R&D pipeline challenge." Chemical biology & drug design 70.6 (2007): 557-565.

4. Grüning, Björn, et al. "Bioconda: sustainable and comprehensive software distribution for the life sciences." Nature methods 15.7 (2018): 475-476.

5. Jumper, John, et al. "AlphaFold 2." (2020).

6. Knight, Steven. "Building software with SCons." Computing in Science & Engineering 7.1 (2005): 79-88.

7. Koch, Linda. "Exploring human genomic diversity with gnomAD." Nature Reviews Genetics 21.8 (2020): 448-448.

8. Landrum, Melissa J., et al. "ClinVar: public archive of interpretations of clinically relevant variants." Nucleic acids research 44.D1 (2016): D862-D868.

9. Li, Biao, et al. "Automated inference of molecular mechanisms of disease from amino acid substitutions." Bioinformatics25.21 (2009): 2744-2750.

10. Lievre, Astrid, et al. "KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer." Cancer research 66.8 (2006): 3992-3995.

11. McCarthy, Davis J., et al. "Choice of transcripts and software has a large effect on variant annotation." Genome medicine 6.3 (2014): 1-16.

12. Pundir, Sangya, et al. "UniProt tools." Current protocols in bioinformatics 53.1 (2016): 1-29.

УДК 579.25

Diversity and properties of bacterial communities associated with White Sea sponges revealed by metagenomics

A. Rusanova[1], D. Sutormin[2], S. Dubiley[1,2]

[1] *Institute of Gene Biology, Russian Academy of Sciences, Moscow, 119334, Russia*
[2] *Skolkovo Institute of Science and Technology, Moscow, 121205, Russia*

Marine sponges (phylum Porifera) and associated microbial communities are classical and at the same time underexplored examples of host-microbial interactions. Depending on the properties of the sponge mesohyl, bacteria can occupy up to 40% of the sponge biomass [1]. These sponge residents can be either symbionts, food for phagocytic sponge cells or pathogens. Sponge-specific symbionts can provide chemical defense, supply the host with essential nutrients and also remove contaminants or metabolic waste products. However, cultivation of the obligate symbionts is a challenging task [2]. In many cases, the only way to investigate their properties are sequencing and comparative genomics. We explored three marine sponges from the Russian Arctic collected in 2016 and 2018: *Isodictya palmata*, *Halichondria panicea* and *Halichondria sitiens*. Analysis of bacterial diversity by 16S rRNA metagenomics (V3-V4 region) revealed the presence of dominant sponge-specific microorganisms for each sponge, which we consider as presumable symbionts. Using shot-gun metagenomics, we recovered metagenome-assembled genomes (MAGs) of these bacteria from metagenome assemblies. Sponge-specific MAG from *H. panicea* was classified as *Amylibacter*, a bacterium which was previously found in the same sponge species from other habitats [3]. Analyses of metabolic features showed the presence of a complete biosynthetic pathway of cobalamin (vitamin B12). Eukaryotes cannot synthesize B12 themselves and have to get it from food or from symbiotic bacteria. B12 biosynthetic pathway was previously found in several bacterial symbionts from different sponges, highlighting it as a symbiotic feature [4]. We performed fluorescence *in situ* hybridization (FISH) experiments to localize the symbiotic bacteria in a sponge tissue. Preliminary data indicates that bacteria are localized in the cell matrix in mesohyl. We also made an attempt of cultivation of homogenized fresh sponge tissue on marine broth plates and obtained small slow-growing colonies, which are to be further classified and characterized.

The work was supported by a grant from the Ministry of Science and Higher Education of Russian Federation (agreement №075-10-2021-114 from 11.10.2021).

References

1. Vacelet, J.; Donadey, C. Electron microscope study of the association between some sponges and bacteria. J. Exp. Mar. Bio. Ecol. 1977, 30, 301–314, doi:10.1016/0022-0981(77)90038-7.

2. Knobloch, S.; Jóhannsson, R.; Marteinsson, V.Þ. Genome analysis of sponge symbiont ‘Candidatus Halichondribacter symbioticus’ shows genomic adaptation to a host-dependent lifestyle. Environ. Microbiol. 2020, 22, 483–498, doi:10.1111/1462-2920.14869.

3. Knobloch, S.; Jóhannsson, R.; Marteinsson, V. Bacterial diversity in the marine sponge Halichondria panicea from Icelandic waters and host-specificity of its dominant symbiont “Candidatus Halichondribacter symbioticus”. FEMS Microbiol. Ecol. 2019, 95, 220, doi:10.1093/FEMSEC/FIY220.

4. Karimi, E.; Keller-Costa, T.; Slaby, B.M.; Cox, C.J.; da Rocha, U.N.; Hentschel, U.; Costa, R. Genomic blueprints of sponge-prokaryote symbiosis are shared by low abundant and cultivatable Alphaproteobacteria. Sci. Rep. 2019, 9, 1–15, doi:10.1038/s41598-019-38737-x.

УДК 577.22

Research of signaling pathways and transcriptional factors activity alteration associated with acute myeloid leukemia

E. Osintseva[1], I. Ruzhenkova[2], E. Belykh[3]

[1] *Higher school of economics, National research university, 109028, 11 Pokrovsky boulevard, Moscow, Russia*
[2] *Russian scientific research institute of hematology and transfusiology, 191024, 16 2d Sovetskaya, Saint Petersburg, Russia*
[3] *BostonGene Corporation, Waltham, Massachusetts, United States*

Acute myeloid leukemia (AML) represents a group of oncohematological neoplasms, which are characterized by uncontrolled proliferation of immature myeloid cells and their accumulation in bone marrow, leading to inhibition of normal hematopoiesis. AML is the most common form of acute leukemia in adults having the shortest survival (5-year survival = 24%) and high rates of relapse as well as the high level of heterogeneity (Shallis et al., 2019). One of the possible reasons for the malignant transformation of hematopoietic cells are mutations, translocations, or aberrant activity of transcriptional factors (TFs) that are involved in normal hematopoiesis maintenance (Khan et al., 2021). Therefore, a better understanding of transcriptional regulation mechanisms can help to develop new therapeutic strategies and to identify promising prognostic markers.

The aim of the current project was to investigate signaling pathways activity alteration and TFs expression in the AML patients cells using RNAseq analysis. Publicly available datasets from Gene Expression Omnibus (GEO) were used: 1) bulk RNA-seq from AML patients and healthy donors (HDs) bone marrow (GSE138702); 2) single cell RNA-seq (scRNA) from AML patients' (GSE116256) and HDs' bone marrow (GSE120221). Also AML dataset with overall survival data was used to perform Kaplan-Meier analysis of survival (dbGaP phs001657.v1.p1).

For this research the tools developed in the Saez lab for the pathway and TF activity investigation were chosen: PROGENy and DoRothEA programs, available as R packages (Garcia-Alonso et al., 2019, Schubert et al., 2018). The authors describe the advantages of PROGENy as the ability to infer the transcriptomic consequences of the processes not by direct mapping of the expression levels of involved genes but by the ‘footprints’ - consistently deregulated genes with the known impact (‘weight’) on the pathway, which also allows to take into account the post-translational protein modifications. In the case of DoRothEA program development several resources of TFs activity estimation were compared and integrated regulons for each TF were derived. The authors proved both these tools outperformed the existing methods being more accurate and informative.

All the tutorials for bulk and pseudo-bulk RNA-seq data analysis were taken from the open GitHub page of Saez Lab (https://github.com/saezlab/transcriptutorial).

At the first stage of the work, we downloaded the bulk RNA sequencing data (GSE138702), preprocessed the data according to the tutorials, and analyzed the activity of signaling pathways using the PROGENy method (detailed description of these steps will be provided below). However, we found a discrepancy with the literature data and subsequently found that there were not enough genes in this dataset for correct further processing (for calculating pathway activity scores). Moreover, one of the AML samples showed extreme scores and could change the whole picture of pathways activity estimation. Thus, we decided to change the dataset.

Therefore, at the next step we downloaded another open source dataset - bone marrow scRNA-seq data of AML patients and HDs. Only untreated AML patients were taken into the analysis. Thus, the work was carried out with 16 AML patients and with 25 HDs. Firstly, scRNAseq data was transformed to pseudo-bulk. Then log2-transformed pseudo-bulk RNAseq counts were visualized using violin plots to choose the threshold for low expressed genes cutoff. We chose the threshold for log2-transformed gene expressions at a level of 1.5. Genes with expression lower than the threshold in less than a half of patients from each group were removed. After low expressed genes removal, 9897 genes remained. Pseudo-bulk RNAseq counts were normalized with the VSN package, in which the variance stabilization and calibration method of normalization was implemented (Huber et al., 2002). After that, principal components analysis (PCA) was carried out and visualized. The obtained plot showed that AML patients and HDs were well separated.

Next, we performed differential expression analysis (DEA) on normalized data using the limma R package (Matthew E. Ritchie et al., 2015). We received a table of genes sorted by p-value, which contains DEA main statistics such as logFC, adjusted p-value. We also plotted qqplot (observed p-value distribution plotted against uniform distribution corresponding to null hypothesis), which demonstrated that the data of healthy and AML samples differed quite a lot.

For estimating the pathway activity, we ran PROGENy (Pathway RespOnsive GENes), a program which allows to infer the pathway activity indirectly, based on consensus gene signatures (Schubert et al., 2018; Holland et al., 2019; Holland et al., 2020). PROGENy was installed as a Bioconductor package. At the first step we downloaded normalized counts and DEA table results obtained in the previous steps to compute PROGENy scores for each sample. The 200 most responsible genes per pathway were chosen for the program running. However, not all of these genes were found in the dataset, which can explain the following possible misaccordance of the results with the literature data. To assess the significance of the pathway activity scores we ran the enrichment analysis and found out the pathways PI3K, MAPK and

VEGF got the highest normalized enrichment scores (NES) and the pathways Trail, EGFR, TNFa got the lowest.

The obtained PROGENy scores for each AML patient and HDs were performed at the heatmap, which illustrated a good separation of pathway activities between two groups. The most activated (PI3K) and deactivated (Trail) pathways correspond to the literature data about pathway activity alteration during AML. PI3K, or, the phosphatidylinositol-3-kinase pathway, is described as important in normal and malignant hematopoiesis, involved in cell proliferation, differentiation and survival (Salihanur Darici et al., 2020). The PI3K pathway is often detected as constitutively activated in AML cells, with FLT3 mutations as one of the driving mechanisms. This matches our dataset's genetic landscape (38% of patients carry FLT3 mutations). Trail is an apoptosis pathway and often deactivated during cancer.

To determine TFs activity alteration in AML compared to healthy controls, we used DoRothEA program (Garcia-Alonso et al., 2019). DoRothEA was installed as a Bioconductor package. After running DoRothEA we obtained 249 TFs whose activity significantly differed in AML patients compared with HDs. As a result, we identified the top 75 most activated and deactivated TFs according to NES values including the previously characterized TFs during AML as well as the new ones.

DoRothEA is a comprehensive resource containing a curated collection of TFs and their transcriptional targets. The set of genes regulated by a specific TF is known as regulon. Thus, we also built the volcano plots for regulon genes of the most differentially activated TFs to identificate the genes of interest for the future deep literature analysis.

At the next step we preprocessed the single cell data (detailed steps are described in the notebook scRNAseq_AML.ipynb at the GitHub repository https://github.com/theCastleBuilder/Acute-myeloid-leukemia) and using the author's markup by cell types looked at the expression of transcription factors in different bone marrow cell types (van Galen et al., 2019). The TFs having the most interesting pattern of expression between cell types were visualized on the dotplot. For instance, *MAFB* gene was shown to be overexpressed preferentially in monocytes and monocyte-like cells which was also displayed at the UMAP plot. Another example is *HOXA9*, which is predominantly expressed in malignant cells (proMonocyte-like, cDC-like, GMP-like).

We began our literature research by studying publications that discuss the relationship between AML and TFs that we found using the DoRothEA package.

For the most activated TF, NCOA1 (nuclear receptor coactivator 1), along with other active NCOA3 and NCOA2, encoded by paralogous genes, no literature evidence about the expression during AML were found. Some authors describe the rearrangements involving these genes resulting in generating the functionally aberrant fusion proteins (S Esteyries et al., 2009). Also for NCOA1 TF the direct interaction

with ASXL1 protein encoded by one of the most frequently mutated genes in malignant myeloid diseases was described (M Katoh, 2013).

The controversial literature data was obtained about the most deactivated TF - FOXP1 - the transcriptional repressor involved in regulation of monocyte differentiation. By Nicolas Duployez et al., the loss-of-function of the *FOXP1* gene was proposed as likely promoting the leukemogenesis mostly in the cases of inv(16)-AML (Duployez et al., 2018). In our dataset there was one patient with inv(16) aberration (van Galen P et al., 2019, Table S1). However, the increased *FOXP1* gene expression was also considered as not a favorable prognostic predictor in cases of intense chemotherapy (Seipel Katja et al., 2020).

One of the interesting chosen TFs was MAFB - a bZip transcriptional factor specific for the monocytic lineage differentiation. This corresponds to our results of MAFB visualization on dot plot and UMAP. Li Yang et al showed the connection between *DNMT3A* R882 mutation and *MAFB* overexpression (Li Yang et al., 2015), which correspond to our dataset genetic properties (44 % of AML patients carried *DNMT3A* mutations). Mutations in *DNMT3A* gene, coding the DNA methyltransferase and responsible for the de novo DNA methylation, considered as driver mutations of AML. This research also demonstrated a positive impact of elevated *MAFB* expression on the overall survival of the patients with *DNMT3A* R882 mutation. Nonetheless, MAFB was described playing a controversial role in leukemogenesis depending on the hematopoietic cells context.

Among the top 100 differentially activated TFs we also selected AHR - ligand-activated transcription factor involved in cellular metabolism, HSC differentiation and immune regulation. There is the literature evidence of the correlation of high *AHR* expression with *FLT3-ITD* mutation, as well as the association with monocytic phenotype (Jennifer N.Saultz at al., 2021). The authors also showed that using the AHR antagonist can facilitate the NK cell mediated killing FLT3-ITD AML cells.

NR4A1 activation can explain the downregulation of MYC and NFkB pathway activities which we observed after PROGENy and DoRothEA run (Salix Boulet et al., 2022).

One of the most activated TF in our research (top 25 by DoRothEA) was MEIS1, a transcriptional regulator participated in hematopoiesis, megakaryocyte lineage development that also can act as cofactor of HOX genes especially in the induction of myeloid leukemias. Both MEIS1 and HOXA9 were activated in our AML samples which conforms to most literature sources about their co-overexpression, associated with poor prognosis (Cailin T. Collins and Jay L. Hess, 2016).

It is worth noting that we did not find any controversial data about MEIS1 activation during AML. That’s why we chose MEIS1 for subsequent Kaplan-Meier

survival analysis. For doing this we took an AML dataset with overall survival data. Patients with NA values were excluded, and finally 123 patients were taken into analysis. Then the patients were divided in two groups depending on the level of MEIS1 expression (the median value was chosen as the threshold).

The Kaplan-Meier curves showed the difference in survival between two groups with decreased survival for the group expressing MEIS1 above median (p-value = 0.005). Thus, MEIS1 can be considered as a possible prognostic marker at the significance level of $\alpha = 0.01$. However, this result should be validated on the bigger AML patient's cohort.

Of course, not for each of the TFs there were references in the literature about its association with AML. Therefore, there is a large space for studying the involvement of not previously mentioned in AML context TFs in gene regulatory networks and signaling pathways in order to build hypotheses. In the future perspective, it will be also interesting to apply the SCENIC framework for the TFs analysis and gene regulatory network (GRN) reconstruction on the scRNA data and compare it with the results we obtained running PROGENy and DoRothEA (Van de Sande et al., 2020).

References

1. Cailin T. Collins and Jay L. Hess. Deregulation of the HOXA9/MEIS1 Axis in Acute Leukemia. Curr Opin Hematol. 2016 Jul; 23(4): 354–361.doi: 10.1097/MOH.0000000000000245

2. Duployez Nicolas, Boudry-Labis Elise, Roumier Christophe et al. SNP-array lesions in core binding factor acute myeloid leukemia. Oncotarget. 2018; 9:6478-6489. https://doi.org/10.18632/oncotarget.24031.

3. Esteyries S, Perot C, Adelaide J et al. NCOA3, a new fusion partner for MOZ/MYST3 in M5 acute myeloid leukemia. Leukemia. v. 22, p. 663–665 (2008)

4. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019 Aug;29(8):1363-1375. doi: 10.1101/gr.240663.118. Epub 2019 Jul 24. Erratum in: Genome Res. 2021 Apr;31(4):745. PMID: 31340985; PMCID: PMC6673718.

5. Holland CH, Szalai B, Saez-Rodriguez J. "Transfer of regulatory knowledge from human to mouse for functional genomics analysis." Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2019. DOI: 10.1016/j.bbagrm.2019.194431.

6. Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, Joughin BA, Stegle O, Lauffenburger DA, Heyn H, Szalai B, Saez-Rodriguez, J. "Robustness and applicability of transcription factor and pathway analysis tools on

single-cell RNA-seq data." Genome Biology. 2020. DOI: 10.1186/s13059-020-1949-z.

7. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002;18 Suppl 1:S96-104. doi: 10.1093/bioinformatics/18.suppl_1.s96. PMID: 12169536.

8. Katoh M. Functional and cancer genomics of ASXL family members. British Journal of Cancer volume 109, p. 299–306 (2013).

9. Khan, I., Eklund, E. E., & Gartel, A. L. (2021). Therapeutic Vulnerabilities of Transcription Factors in AML. Molecular cancer therapeutics, 20(2), 229–237. https://doi.org/10.1158/1535-7163.MCT-20-0115

10. Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 2015, Vol. 43, No. 7. doi: 10.1093/nar/gkv007

11. Salihanur Darici, Hazem Alkhaldi, Gillian Horne, et al. Targeting PI3K/Akt/mTOR in AML: Rationale and Clinical Evidence. J Clin Med. 2020 Sep; 9(9): 2934. doi: 10.3390/jcm9092934

12. Saultz Jennifer N. et al. Aryl Hydrocarbon Receptor (AHR) Gene Expression in AML Is Associated with FLT3-ITD+ AML and HLA-E Induced Immune Resistance Reversed By Ik-364. Blood (2021) 138 (Supplement 1): 1161.https://doi.org/10.1182/blood-2021-147589.

13. Schubert, M., Klinger, B., Klünemann, M. et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun 9, 20 (2018). https://doi.org/10.1038/s41467-017-02391-6

14. Seipel Katja, Messerli Christian, Wiedemann Gertrud et al. MN1, FOXP1 and hsa-miR-181a-5p as prognostic markers in acute myeloid leukemia patients treated with intensive induction chemotherapy and autologous stem cell transplantation. Leukemia Research, 89(), 106296–. doi:10.1016/j.leukres.2020.106296

15. Shallis RM, Wang R, Davidoff A, Ma X, Zeidan AM. Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. Blood Rev. 2019 Jul;36:70-87. doi: 10.1016/j.blre.2019.04.005. Epub 2019 Apr 29. PMID: 31101526.

16. Van de Sande, B., Flerin, C., Davie, K. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nat Protoc 15, 2247–2276 (2020). https://doi.org/10.1038/s41596-020-0336-2

17. Van Galen P, Hovestadt V, Wadsworth Ii MH, et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell. 2019;176(6):1265-1281.e24. doi:10.1016/j.cell.2019.01.031

18. Yang Li, Liu Ya'Nan, Zhu Li et al. DNMT3A R882 mutation is associated with elevated expression of MAFB and M4/M5 immunophenotype of acute myeloid leukemia blasts. Leuk Lymphoma. 2015;56(10):2914-22. doi: 10.3109/10428194.2015.1015123.

УДК 577.22

Genetic variant annotation in introns branchpoints

A. Sergeeva[1], I. Veretenenko[1], M. Skoblov[2], Y. Barbitoff[1]

[1] *Bioinformatics Institute, St. Petersburg, Russia*
[2] *Research Center for Medical Genetics, Moscow, Russia*

A key step in the molecular diagnosis of hereditary diseases is the interpretation of genetic variants found during genome or exome sequencing. At the moment, however, interpretation is quite a challenge, and the cumulative efficiency of molecular diagnostics ranges from 30 to 50%. This is partly due to the lack of good methods for annotating different classes of genetic variants that do not involve changes of the amino acid sequence of a protein.

This work is a part of a large project aimed at enhancing genetic variant annotation. This part is devoted to predicting effects of genetic variants on intra-intron sequences involved in splicing (primarily, in the region of the branchpoint). The aim of the work is to determine the effects (like pathogenicity) of variants located in branchpoints.

In the course of the study, the project was divided into two main parts: variant annotation based on existing annotators and new predictive model development.

The first part of the current study was based on the previous paper which evaluated the prediction accuracy for branchpoint-predicting tools [1]. According to this paper, the highest accuracy was shown by the combination of two predictors: Branchpointer [2] and BPP [3]. We used both of them in order to predict the branchpoints of introns in two databases: ClinVar and gnomAD. The ClinVar database provides information about the clinical impact of variation while the gnomAD database described the natural variation in healthy individuals. We suggested that variants found in gnomAD should have lower impact on branchpoint efficiency compared to pathogenic variants from ClinVar. Prior to analysis, we removed indels from both databases as indels change genomic coordinates of the analyzed regions.

We found out that Branchpointer was significantly slower and memory consuming than BPP, therefore it might be inconvenient to use Branchpointer with a large dataset. However, we managed to analyze the separate impact of 9,240 ClinVar and 208,550,834 gnomAD SNPs on the branchpoint prediction via both BPP and Branchpointer.

Both algorithms have shown the significant difference between branchpoint scores (or probabilities) of ClinVar and gnomAD databases (p-value < 0.05 in Wilcoxon-Mann-Whitney test) which demonstrated the study importance.

Expectedly, we discovered that ClinVar mutations more often showed the reference branchpoint score reducing according to both predictors. At the same time, we found certain ambiguity of predictions by BPP and Branchpointer which did not allow us to draw confident conclusions regarding the magnitude of the differences.

Thus, we realized that it is necessary to create our own branchpoint predictor. To do this, we decided to use the ML architecture. High-confidence branchpoints [4] were used as training data, on which the Branchpointer was also trained. The main idea of the model is to use k-mers and distance to 3' ends to predict the position of branchpoints. Our implementation is motivated by the typical location and structure of branchpoint regions that includes polypyrimidine tract. The model has two main components:

First, we calculate branchpoint probabilities for all intron position based on k-mers in (position-n, position+n) area. The probabilities are calculated by a pre-trained model. To train the model, we used both positive and negative examples of branchpoints. Experiment-based high confidence branchpoints were used as positive examples, whereas as negative examples included random positions from introns that are not high confidence branchpoints.

The parameters *k, n* and the classification method (lightgbm, catboost, xgboost, random forest, naive bayes and SVM) were chosen by selecting the highest value of AUC and accuracy. Notably, the prediction accuracy did not depend strongly on *n* but was strongly affected by changes in the value of *k*. We chose $n = 70$, $k = 5$ and Random Forest Classifier as showing the optimal results.

The accuracy of the constructed model was much better compared to other predictors [2]. The average score for all chromosomes was 99.75 for our method versus 94.73 for for branchpointer, 88.8 for Naive Bayes, 82.8 for HSF and 74.9 for SVM-BPFinder.

After obtaining per-base branchpoint probabilities, positions of branchpoints were selected within the last 50 positions of every intron by choosing the one with the largest probability of branchpoint. Exact matches to known locations of high-confidence BPs were obtained in only 25% of cases, 73% of cases were in the vicinity of 5 from the known position. Thus, we have obtained a model that can determine the positions of branchpoints with good accuracy, but new functions can be added to further improve it.

Therefore, our future plan is to develop an annotator similar to BBP and Branchpointer that will predict the pathogenicity of all variants in introns. This stage would require further development of the constructed model. We also plan to compare predictions of our model with BBP and Branchpointer and eventually create a common tool that most reliably predicts the effects of genetic variants. We hope that we will be able to create the most accurate branchpoint annotator that includes the best of the three models.

References

1. Leman R, Tubeuf H, Raad S, Tournier I, Derambure C, Lanos R, Gaildrat P, Castelain G, Hauchard J, Killian A, Baert-Desurmont S, Legros A, Goardon N, Quesnelle C, Ricou A, Castera L, Vaur D, Le Gac G, Ka C, Fichou Y, Bonnet-Dorion F, Sevenet N, Guillaud-Bataille M, Boutry-Kryza N, Schultz I, Caux-Moncoutier V, Rossing M, Walker LC, Spurdle AB, Houdayer C, Martins A, Krieger S. Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants. BMC Genomics. 2020 Jan 28;21(1):86. doi: 10.1186/s12864-020-6484-5. PMID: 31992191; PMCID: PMC6988378

2. Signal, B., Gloss, B. S., Dinger, M. E., & Mercer, T. R. (2018). Machine learning annotation of human branchpoints. Bioinformatics (Oxford, England), 34(6), 920–927. https://doi.org/10.1093/bioinformatics/btx688

3. Zhang Q, Fan X, Wang Y, Sun MA, Shao J, Guo D. BPP: a sequence-based algorithm for branch point prediction. Bioinformatics. 2017 Oct 15;33(20):3166-3172. doi: 10.1093/bioinformatics/btx401. PMID: 28633445

4. Mercer, T. R., Clark, M. B., Andersen, S. B., Brunck, M. E., Haerty, W., Crawford, J., Taft, R. J., Nielsen, L. K., Dinger, M. E., & Mattick, J. S. (2015). Genome-wide discovery of human splicing branchpoints. Genome research, 25(2), 290–303. https://doi.org/10.1101/gr.182899.114

УДК 579.25

Analysis of RecQ involvement in primed adaptation in the type I-E CRISPR-Cas system of *Escherichia coli*

A. Shiriaeva[1,2], Y. Tsoy[3], K. Severinov[1,4]

[1] *Skolkovo Institute of Science and Technology, Moscow 121205, Russia*
[2] *Saint Petersburg State University, Saint Petersburg, 199034, Russia*
[3] *Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, 195251, Russia*
[4] *Waksman Institute, Rutgers, State University of New Jersey, Piscataway, 08854, USA*

The type I-E CRISPR-Cas system of *Escherichia coli* includes a CRISPR array and *cas* genes (Brouns et al., 2008). The CRISPR array is composed of 28-bp DNA repeats separated by unique 33-bp spacers. Spacers acquired from phage genomes provide bacteria with immunity against bacteriophages (Barrangou et al., 2007). The immunity is achieved by the cleavage of phage sequences complementary to spacers (protospacers) by the Cas3 protein (Westra et al., 2012). Due to its helicase and endonuclease activities, Cas3 cleaves DNA around the protospacer into fragments of unknown size. Other cellular nucleases further degrade these fragments (Kurilovich et al., 2019).

The protospacer cleavage leads to acquisition of new spacers from the protospacer-flanking regions during a process called primed adaptation (Datsenko et al., 2012). The first step of spacer acquisition is excision of a short spacer precursor called a prespacer (Shiriaeva et al., 2019). The prespacer is partially double-stranded and consists of a 33-nt strand and a 37-nt strand. It is currently not clear how prespacers are generated. Based on our previous findings we suggested a model where a complex of two Cas proteins, Cas1 and Cas2, binds to a sequence of a future spacer within a long DNA molecule and protects it from degradation by Cas3 and other nucleases. The unprotected ends are trimmed by single-strand specific nucleases, such as the 5'-3' RecJ exonuclease. If the ends are double-stranded, the RecBCD helicase/nuclease unwinds them to provide access to the 5' ends for RecJ. Surprisingly, only a 3-fold decrease in spacer acquisition efficiency is observed in Δ*recB* cells. We suggest that another helicase, RecQ, substitutes for the RecBCD helicase activity during prespacer generation in Δ*recB* cells.

To test this hypothesis, we studied primed adaptation in *wt*, Δ*recQ,* Δ*recB,* and Δ*recB* Δ*recQ* strains. Since new spacers are integrated into the beginning of CRISPR arrays, amplification of this region using PCR gives products of various lengths. The shortest product corresponds to initial, nonexpanded CRISPR arrays. The longer products correspond to CRISPR arrays with one or several newly acquired spacers. Using high-throughput sequencing of resulting amplicons, spacer acquisition efficiency can be calculated as the number of newly acquired spacers divided by the total amount of sequenced CRISPR arrays. Six biological replicates of the primed adaptation experiment were performed for each of the four strains, and spacer

acquisition efficiency was calculated based on the sequencing data. Levene's test showed that variances are not equal. Therefore, Welch's t-test was applied to compare means between the pairs of strains, and p-values were adjusted using Benjamini and Hochberg's p-adjustment method. The results confirm a 3-fold decrease in primed adaptation efficiency in the Δ*recB* mutant compared with the *wt* (p_{adj}=0.00005). No significant differences were observed between the *wt* and Δ*recQ* cells (p_{adj}=0.07). However, a 7-fold decrease in spacer acquisition efficiency was observed in Δ*recB* Δ*recQ* compared to Δ*recB* cells (p_{adj}=0.003). This result supports our hypothesis that RecQ is involved in primed adaptation in the absence of RecBCD, though further *in vitro* and *in vivo* studies are required to elucidate the exact role of RecQ in this process.

References

1. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science 315, 1709–1712. https://doi.org/10.1126/science.1138140.

2. Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321, 960–964. https://doi.org/10.1126/science.1159689.

3. Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K., and Semenova, E. (2012). Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun 3, 945. https://doi.org/10.1038/ncomms1937.

4. Kurilovich, E., Shiriaeva, A., Metlitskaya, A., Morozova, N., Ivancic-Bace, I., Severinov, K., and Savitskaya, E. (2019). Genome Maintenance Proteins Modulate Autoimmunity Mediated Primed Adaptation by the Escherichia coli Type I-E CRISPR-Cas System. Genes 10, 872. https://doi.org/10.3390/genes10110872.

5. Shiriaeva, A.A., Savitskaya, E., Datsenko, K.A., Vvedenskaya, I.O., Fedorova, I., Morozova, N., Metlitskaya, A., Sabantsev, A., Nickels, B.E., Severinov, K., et al. (2019). Detection of spacer precursors formed in vivo during primed CRISPR adaptation. Nat Commun 10, 4603. https://doi.org/10.1038/s41467-019-12417-w.

6. Westra, E.R., van Erp, P.B.G., Künne, T., Wong, S.P., Staals, R.H.J., Seegers, C.L.C., Bollen, S., Jore, M.M., Semenova, E., Severinov, K., et al. (2012). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol. Cell 46, 595–605. https://doi.org/10.1016/j.molcel.2012.03.018.

УДК 577.22, 004.94

Dissecting the role of gene expression variability in complex traits

M. Slizen[1], Y. Barbitoff[2]

[1] *Institute of Protein Research, Institutskaya str. 4, Pushchino, Moscow Oblast, 142290, Russia*
[2] *Bioinformatics Institute, Kantemirovskaya str. b.2a, 197342 Saint Petersburg, Russia*
Email: mikha.shtol@gmail.com

Genome-Wide Association Study (GWAS) is a technique used to look for genome sequence variations that affect the development of complex traits. In recent years, GWAS results have been published for thousands of different traits, including two of the world's largest datasets, UK Biobank and Finngen. It is known that changes in gene expression levels are one of the main mechanisms that determine the small effects of genetic variants detected during GWAS. In this project, we sought to test the hypothesis that not only the level of gene expression, but also the degree of expression variability, is associated with the influence of a gene on complex human traits.

GWAS for pulse rate from UKBB was used as a test dataset. LD-based clumping was performed to identify lead SNPs with p-value < 10^{-8}, with LD information taken from HapMap2 data. Genes closest to lead SNPs were identified by intersection with GRCh37 RefSeq Reference Genome Annotation gff3 file. GTEx Analysis V8 was used to obtain gene expression data. For each gene in each of the 54 tissues we calculated ranges of expression normalized by median value. Then, inter-tissue mean and maximum ranges of variability expression were calculated.

Among 56200 genes from GTEx we selected 44025 genes with non-zero expression at least in one tissue. 329 lead SNPs identified in the pulse rate GWAS were annotated with 363 closest genes. After deduplication and retrieval of variability metrics for genes with non-zero expression presented in GTEx, 87 genes were left. . Distribution of expression variability metrics for this subset and a total set of GTEx genes were compared by t-test for two distributions with unequal variance. Comparison of 3 metrics out of 4 (mean ranges, mean "centered" ranges and maximal ranges) have showed significant differences (p-value<0.01) and one metric (maximal "centered" ranges) have showed less significant difference (p-value=0.06). Therefore, we can conclude that genes at GWAS loci tend to have a wider average range of expression. In future research we plan to refine our statistical approach and test our assumption for multiple GWAS datasets.

More details can be found in GitHub repository https://github.com/MShtol/expression_variability

УДК 575.1, 575.85

Determining the effectiveness of momi2 for inferring demographic history in GADMA

K. Struikhina[1,4], V. Mikhalchuk[2], E. Noskova[3]

[1] *Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia*
[2] *Peter the Great St. Petersburg Polytechnic University, Polytechnicheskaya st. 29, 195251, St. Petersburg, Russia*
[3] *ITMO University, Kronverksky pr. 49A, 197101, St. Petersburg, Russia*
[4] *Higher School of Economics, Myasnitskaya st. 20, 101000, Moscow, Russia*

GADMA implements methods for automatic inference of the joint demographic history of multiple populations from genetic data. The demographic history of populations is the history of their development with parameters such as the size of populations, the time of their divergence, the rate of migration and selection. Based on genetic data, these parameters can be reconstructed using various statistical methods.

GADMA is based on the three open-source packages for inferring demographic history: dadi, moments, and momi2. This project determines the accuracy of demographic history output using the momi2 engine added to GADMA. Our aim was to determine the effectiveness of momi2 for inferring demographic history in GADMA.

We simulated genetic data for the selected history, then restored parameters using GADMA with momi2 engine (for one chromosome and the whole genome) and compared the obtained parameters with the original ones.

As a result, momi2 determined the parameters of the demographic population on the entire genome better than on one chromosome. However, the epoch time parameter still does not correspond to the real value.

УДК 004.94, 57.08

Benchmark creation for drug-target interaction (DTI) prediction task

D. Traktirov[1], E. Kartysheva[2]

[1] *I.P. Pavlov Department of Physiology, Federal State Budget Scientific Institution "Institute of Experimental Medicine", St. Petersburg, 197376, Russia.*
[2] *JetBrains Limited, 4, Pindou, Egkomi, 2409 Nicosia, Cyprus*

Introduction

Drug-target interaction prediction (DTI) task plays an important role in the drug discovery process, which aims to identify new drugs for biological targets. Automation of prediction will speed up the process of creating new drugs. There are many machine learning models that solve this problem [3, 5, 7] nowadays, however, due to the presence of a huge number of different datasets and testing protocols, it is difficult to compare different models with each other. And so one unified benchmark is needed.

The aim of this project was to create a benchmark for drug-target interaction (DTI) prediction task. In order to achieve the aim we needed to select suitable datasets to perform pipeline (potential candidates are KiBA, Yamanishi_08, Davis), create an evaluation protocol, and finally implement several most relevant models and test them using the created evaluation protocol.

Materials and methods

Benchmark was implemented using Python (version 3.7.13) libraries PyTorch [6], DGL [11] and Sckit-Learn [12].

BindingDB and Davis are chosen as two main datasets [1, 4]. BindingDB and Davis are both databases of measured binding affinities, focusing chiefly on the interactions of proteins considered to be candidate drug-targets with ligands that are small, drug-like molecules. As of May 4, 2022, BindingDB contains 41,296 Entries, each with a DOI, containing 2,513,948 binding data for 8,839 protein targets and 1,077,922 small molecules. Davis Kinase binding affinity dataset contains the interaction of 72 kinase inhibitors with 442 kinases covering >80% of the human catalytic protein kinome.

DistMult, TriModel and KGE_NFM are chosen as three baseline models. DistMult is a knowledge graph embedding (KGE) model that allows to learn the low-rank representations for all entities and relations using one embedding vector [8]. In terms of current task entities are drugs and proteins, whereas relation is the existence of interaction between them. TriModel is a KGE model based on tensor factorization that extends the DistMult and ComplEx models [10]. It represents each

entity and relation using three embedding vectors. Neural Factorization Machine (NFM) combines the linearity of FM in modeling second-order feature interactions and the non-linearity of neural network in modeling higher-order feature interactions [7]. KGE_NFM model allows one to train one of the KGE models and pass its embeddings into the NFM model as features (along with drugs and target features).

The code is stored at https://github.com/Lemonnik/BI_2021_JB_benchmark

Results

In the current work two baseline datasets, Davis and BindingDB, and three baseline models were implemented. Python script that uses our evaluation protocol was created. The script allows the user to train and test one of the implemented models on one of the implemented datasets, or use his own model/dataset. As a result of the script, the user gets evaluation metrics, such as AUROC, AUPRC, precision and recall, in order to quantify the performance of a predictive model. We hope that the proposed benchmark will standardize the model testing process in the future.

References

1. Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Nature biotechnology 2011,29, 1046–1051

2. Dunham, Brandan, and Madhavi K. Ganapathiraju. 2022. "Benchmark Evaluation of Protein–Protein Interaction Prediction Algorithms" Molecules 27, no. 1: 41. https://doi.org/10.3390/molecules27010041

3. Kexin Huang, Cao Xiao, Lucas M Glass, Jimeng Sun, MolTrans: Molecular Interaction Transformer for drug–target interaction prediction, Bioinformatics, Volume 37, Issue 6, 15 March 2021, Pages 830–836, https://doi.org/10.1093/bioinformatics/btaa880

4. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 2007;35(Database issue):D198-D201. doi:10.1093/nar/gkl999

5. Sameh K Mohamed, Vít Nováček, Aayah Nounu, Discovering protein drug targets using knowledge graph embeddings, Bioinformatics, Volume 36, Issue 2, 15 January 2020, Pages 603–610, https://doi.org/10.1093/bioinformatics/btz600

6. Paszke A. et al. (2019) Automatic differentiation in Pytorch. In: NeurIPS, Vancouver, Canada.

7. Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics. In Proceedings of the 40th International ACM SIGIR

Conference on Research and Development in Information Retrieval (SIGIR '17). Association for Computing Machinery, New York, NY, USA, 355–364. https://doi.org/10.1145/3077136.3080777

8. Yang, Bishan & Yih, Wen-tau & He, Xiaodong & Gao, Jianfeng & Deng, li. (2014). Embedding Entities and Relations for Learning and Inference in Knowledge Bases.

9. Ye, Q., Hsieh, CY., Yang, Z. et al. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. Nat Commun 12, 6775 (2021). https://doi.org/10.1038/s41467-021-27137-3

10. Sameh K Mohamed, Vít Nováček, Aayah Nounu, Discovering protein drug targets using knowledge graph embeddings, Bioinformatics, Volume 36, Issue 2, 15 January 2020, Pages 603–610, https://doi.org/10.1093/bioinformatics/btz600

11. Wang et. al., Deep graph library: A graph-centric, highly-performant package for graph neural networks, arXiv, 2019

12. Pedregosa et. al., Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.

УДК 004.94, 577.22

Clustering Hi-C contact graphs using Graph Neural Networks

F. Velikonivtsev[1,2], I. Tolstoganov[3], A. Korobeynikov[3]

[1] *Almazov National Medical Research Center, Akkuratova str. 2, 197341, Saint-Petersburg, Russian Federation*
[2] *Bioinformatics Institute, Kantemirovskaya str. b.2a, 197342, Saint Petersburg, Russia*
[3] *Center for Algorithmic Biotechnology, St. Petersburg State University, 10 liniya, Vasilievsky island, 11d, 199004, Saint-Petersburg, Russian Federation*

Introduction

Metagenomic binning is a problem of restoring microbial species from the given metagenomic samples by grouping assembled contigs together according to some similarity measure among contained contigs and therefore obtaining bins with metagenome assembled genomes (MAGs). Some state-of-the-art technologies from deep learning, such as variational autoencoders, are used in these tools. For instance, VAMB (Variational Autoencoder for Metagenomic Binning) is a binning tool that allows to embed contig profiles - combined data from TNF and coverage profiles (data with tetranucleotides frequencies and coverage metrics per contig) - in the latent feature space and thus perform good-quality clustering [1].

However, despite being able to restore high-quality genomes (genomes with >95% completeness and <5% contamination), current binning strategies are still not perfect: if obtained from a single sample, binning results can be erroneous, and additionally, as similarity measure derives from contig profile, it is hard for these tools to distinguish closely related species and strains present in the sample. Thus, it can be not enough to rely only on contig profiles.

Hi-C method is the approach aimed at reconstructing chromosomal structure within a genome [2]. It allows to capture genome regions in close spatial proximity which could have been far in the strand and sequence them together. This technology is widely used for exploring three-dimensional genome interaction and architecture, and moreover, it can be used as a starting point for chromosome and genome 3-D structure reconstruction [3]. Thus, applying fundamentals of method to binning problem, we suppose that pair of contigs from the same genome can possess significantly more Hi-C links between each other than pair of contigs from separate genomes.

Hi-C contact map is a set of pairwise Hi-C interactions between contigs, and its usage provides us with several benefits. It is clear that a contact map can lead to additional contigs structure reveal. Contact map`s natural representation is a weighted graph, where nodes are contigs and edges represent the number of Hi-C contacts between a pair of contigs. Thus, graph clustering becomes equal to the metagenomic

binning problem, and graph deep learning methods, such as graph neural networks (GNNs), can be naturally applied for this task, though the binning tool that does not utilize deep learning methods – bin3C – has already been developed and showed good performance [4]. Next big advantage of utilizing Hi-C contact map can be independence from the number of samples – contact map can be obtained from single sample data and already provide accurate and complete information about contigs colocalization.

Thereby, in our work we aimed to explore opportunities of clustering contig contact maps using graph neural networks. For our project we have chosen to work with 2 recently developed GNN models – DMoN (Deep Modularity Network) [5] and GraphMB [6]. DMoN utilizes modularity maximization as a loss-function in learning and requires prepared features and doesn`t support edge features, while GraphMB uses modified GraphSAGE as a working net and also provides autoencoder from previously mentioned VAMB for efficient features extraction, but supports edge features as well.

We have created binning pipeline involving data preparation, clustering, and binning results evaluation. We chose VAMB`s performance as baseline as it showed its efficiency, and moreover, we compared GNN`s binning results with bin3C results where it was possible. We ran this pipeline on 3 datasets – Zymo, IC9 and synthetic CAMI AIRWAYS [8] – which differ by size and graph sparsity. Here we present our work protocol and current outcomes.

Materials and methods

We used 3 input datasets with different assembly size and presense of golden standard:

1. Zymo – supervised, 10 genomes (size from 2 to 27 mbp), 6625 contigs, 76799 Hi-C links;

2. IC9 – unsupervised, approximately 16 genomes, 165712 contigs, 1150887 Hi-C links;

3. Synthetic CAMI AIRWAYS (was simulated by Sim3C [7] from CAMI) – unsupervised, 600 genomes (size from 500 bp to 3 mbp), 728682 contigs, 70405 Hi-C links.

These datasets included: assembled contigs in *.fasta* format, contact map in *.tsv* format, contig depth information in tab-separated format, and ground truth for contigs with known MAG belongings (for supervised datasets) in *.tsv* format.

We used following tools for clustering contact map: i) YAMB (v. 3.0.3) on all datasets; ii) GraphMB (v. 0.1.3) on all datasets; iii) DMoN (v. 0.1) on Zymo dataset; iv) Bin3C (v. 0.1.1) on Zymo dataset.

Binning results were evaluated by following tools:

1. AMBER (v. 2.0.3) – for supervised datasets (CAMI AIRWAYS, Zymo);

2. CheckM (v. 1.1.3) – for unsupervised IC9 dataset.

Binners` output were transformed to proper AMBER format for Zymo and CAMI AIRWAYS datasets.

Main metrics comprised the tools comparison were: completeness, contamination, purity of bins and number of restored high-quality genomes (HQ genomes – bins with >90% completeness, <5% contamination).

For CAMI AIRWAYS dataset, VAMB and GraphMB was run with minimal contig lengths 200, 500, 1000, 2000. For Zymo dataset, GraphMB was tested with default parameters in first case and with additional parameter allowing to obtain additional features for latent feature transformation – TNF profiles. Input contact map and feature matrix for DMoN were written to *.npz* format as matrix of features and adjacency, for GraphMB input contact map was transformed and written in *.gfa* format. Hi-C score values for edge features were passed log-scaled and squareroot-scaled.

While GraphMB was obtaining contig features on a run, DMoN required them already prepared, thus, we tried TNF profiles of contigs, latent features from VAMB work as contig features and empty feature vector to check whether it is capable to catch information only from adjacency matrix.

Additionally, we performed sanity-check for DMoN and ran it on synthetic non-DNA graph consisted of 10 disjoint connected components-cliques each of size 10 and according features as one-hot encoding of clique belonging – it was done because DMoN showed bad performance on Zymo dataset (see Results section).

Performed pipeline steps, wide instructions for each one and tools requirements are available at GitHub pipeline page (can be accessed through link).

Results

DMoN performance was evaluated by counting number of restored HQ genomes. It turned out that DMoN showed bad performance as it restored 0/10 HQ genomes in Zymo dataset in all 3 scenarios (latent features, TNF profiles, empty features). Results stood the same on a row of repetitive runs. The requirement of sanity-check was clear, DMoN was tested on synthetic graph with unambiguous features – only 8/10 clusters were restored correctly, erroneous bin had 50% purity – entire non-overlapping cluster was put in that bin. According sanity-check results, we considered these simple sanity-check as failed, DMoN was taken out of further experiment, work continued for GraphMB.

On Zymo dataset GraphMB restored 7/10 HQ genomes without using TNF profiles and 8/10 with usage of latter whilst VAMB restored 10/10 and bin3C restored 6/10. On CAMI AIRWAYS dataset GraphMB restored 98/600 HQ genomes vs VAMB`s 93/600 HQ genomes. On IC9 dataset GraphMB restored 12 HQ genomes as well as VAMB restored 12. No significant differences between log-scaling and squareroot-scaling of Hi-C score was observed.

Discussion

Insufficient performance of DMoN on given data can be explained by 2 things – firstly, it turns out that DMoN strongly relies on given features and leaves the adjacency second priority (failed sanity-check indirectly proves it, as proposed test graph was clearly splittable on 10 clusters only by looking on adjacency). It also clean from the formula in paper [5] that features must nonempty as they take part in learning and gradient step, thus, absent features, or improper features lead to died gradients and stop of a training process. Secondly, perhaps developed for broader usage and not directly for bioinformatics problems, DMoN postulates flexibility in the clusterization problem but lacks particular solutions for metagenomic binning problem, where only single modularity metrics is not enough for correct clustering.

GraphMB, as well as VAMB, doesn`t possess the latter drawback – they were initially developed for metagenomic binning task. While VAMB successfully generates informative features and bins contigs relying on features only, GraphMB tries to improve binning by relating contigs between each other through Hi-C contact map and then by clustering node embeddings. It`s worth noting that by default GraphMB uses VAMB to generate latent features from contigs, thereby it strongly relies on VAMB and its similar performance is explained by this influence in some extent. As can be seen on CAMI AIRWAYS dataset (which is large enough to make VAMB binning less successful then GraphMB binning) where GraphMB restored more HQ genomes then VAMB, contact map indeed provides additional information, crucial for further clustering, this supports our suggestion about competitive results of GraphMB, though it restored less bins on Zymo dataset then VAMB did. It can be explained by the nature of Hi-C method: it captures chromosome interactions within single cell, that`s why accordance on Hi-C contact map can lead to binning distinct chromosomes of an organism to distinct bins and not to a single bin containing whole genome. And it turned out that Zymo dataset does indeed have genome of yeast *S. cerevisiae,* which has 17 chromosomes, and GraphMB clustered up to 37% of its genome to single bin and left other ration unclustered, whereas VAMB succeeded with this genome. It can show us that Hi-C contact map can provide both neat information for binning and can potentially cause misclustering in the case of bad Hi-C coverage between separated parts of a genome like chromosomes.

To summarize, Hi-C contact map is a good alternative for metagenomic binning problem, it captures spatial integration information of contigs and can further been clustered by graph neural networks. One of our tested GNNs – GraphMB – showed

good performance according to the baseline and even run better on large dataset. Further study should be driven to investigating contact map implicit properties.

References

1. Nissen, J.N., Johansen, J., Allesøe, R.L. et al. Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol 39, 555–560 (2021). https://doi.org/10.1038/s41587-020-00777-4

2. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*. 2012;58(3):268-276. doi:10.1016/j.ymeth.2012.05.001

3. van Berkum NL, Lieberman-Aiden E, Williams L, et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*. 2010;(39):1869. Published 2010 May 6. doi:10.3791/1869

4. DeMaere, M., Darling, A. bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. Genome Biol 20, 46 (2019). https://doi.org/10.1186/s13059-019-1643-1

5. Tsitsulin, Anton & Palowitch, John & Perozzi, Bryan & Müller, Emmanuel. (2020). Graph Clustering with Graph Neural Networks.

6. Metagenomic binning with assembly graph embeddings. Andre Lamurias, Mantas Sereika, Mads Albertsen, Katja Hose, Thomas Dyhre Nielsen. bioRxiv 2022.02.25.481923; doi: https://doi.org/10.1101/2022.02.25.481923

7. Matthew Z DeMaere, Aaron E Darling, Sim3C: simulation of Hi-C and Meta3C proximity ligation sequencing technologies, *GigaScience*, Volume 7, Issue 2, February 2018, gix103, https://doi.org/10.1093/gigascience/gix103

8. Fritz, A., Hofmann, P., Majda, S. et al. CAMISIM: simulating metagenomes and microbial communities. *Microbiome* 7, 17 (2019). https://doi.org/10.1186/s40168-019-0633-6

УДК 579.25, 579.8

Systematics and classification of plasmids

E. Vostokova, *P*. Vychyk, M. Rayko

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

Plasmids are extrachromosomal DNA molecules, predominantly circular in shape, capable of autonomous replication. Although plasmids are an optional part of bacterial genomes, their acquisition by a cell often provides significant adaptive advantages for the host cell through the acquisition of new genes providing antibiotic resistance, the ability to utilize new substrates or the expression of virulence factors. Plasmids are also significant as one of the main drivers of horizontal gene transfer in bacterial evolution. There are various approaches for the classification of plasmids: based on the phenotype of the host cell, conjugative transfer systems used by plasmids, and the ability to replicate in various taxonomic groups.

The aim of our work was to create an approach to the classification of plasmids, which could be used further to determine the relationship of plasmids first obtained in metagenomic projects. The approach we considered is based on the clustering of Rep protein sequences involved in the replication of the vast majority of circular plasmids.

The main sources of data for plasmids protein sequences were NCBI RefSeq plasmid protein database and unpublished data from the project supervisor.

Rep proteins inference in the data sequences was performed with hidden Markov models for Rep_1, Rep_2, Rep_3, Duff1424 families available in PFAM. Sequences of Rep-proteins were obtained from plasmids proteome data with hmmsearch from HMMER 3.3. The gained proteins were clustered at 50% similarity using MMseqs2 to decrease the redundancy. Derived sequences were further aligned with Mafft aligner tool v.7.453. Resulted Rep-proteins sequence alignment was used to build phylogeny with FastTree 2.1.11 and IQ-TREE 1.6.12. Another method for studying protein sequence relationships was graph network reconstruction with Gephi 0.9.2. The step for tree reconstruction was automatized with the script gettree.py (available in the repository).

As a result, 5288 proteins with Rep-like domain were identified in NCBI GenBank plasmid proteome (over 1 500 000 non-redundant sequences) and supervisor's sequencing data using HMM-models from Pfam. Clustering at 50% sequence identity level reduced combined dataset size to 1255 proteins, which were further used for reconstruction of graph and phylogeny.

Our results show that Rep-proteins contain only one type of the Rep-domain. The most numerous types were the Rep_3 family (918 proteins out of 1255 total). Acquired phylogenetic tree and graph topology demonstrate sub-groups existence within Rep_3 family and sufficient level of credibility.

References

1. Brooks L, Kaze M, Sistrom M. A Curated, Comprehensive Database of Plasmid Sequences. Microbiol Resour Announc. 2019;8:e01325-18.

2. Kirstahler P, Teudt F, Otani S, Aarestrup FM, Pamp SJ. A Peek into the Plasmidome of Global Sewage. mSystems. 2021;6:e00283-21.

3. Shintani M, Sanchez ZK, Kimbara K. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. Front Microbiol. 2015;6.

4. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. Mobility of Plasmids. Microbiol Mol Biol Rev. 2010;74:434–52.

5. del Solar G, Giraldo R, Ruiz-Echevarría MJ, Espinosa M, Díaz-Orejas R. Replication and Control of Circular Bacterial Plasmids. Microbiol Mol Biol Rev. 1998;62:434–64.

УДК 577.22, 57.08

Differential expression analysis of macrophage RNA sequencing data using the Hobotnica tool

A. Zhelonkin[1], A. Belyaeva[2], E. Karpulevich[3]

[1] *Saint-Petersburg State University, 7-9 Universitetskaya Embankment, St Petersburg, Russia*
[2] *Lomonosov Moscow State University, Leninskie Gory, Moscow*
[3] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25 Alexander Solzhenitsyn st, Moscow*

High-throughput RNA-seq is widely used in the differential expression (DE) analysis. Differentially expressed genes are key to understanding phenotypic variation. Wide range of tools have been developed for DE analysis [1]. Each has its own specificities in the assumptions on the statistical properties inherent to RNA-seq data [2]. Consensus hasn`t been reached as to the best pipeline for correctly identifying differentially expressed genes from RNA-seq data.

Hobotnica is a tool that was developed to put clarity into the choice of the best DE tool for a given RNA-seq dataset. Hobotnica computes and visualizes gene-level RNA-seq differential expression between two experimental conditions by 5 tools with different statistical methods embedded in them: DESeq, EBSeq, edgeR, limma voom and NOISeq. Hobotnica also compares results and calculates the best differential expression tool for input data based on a H-score summary statistic, simultaneously visualizing intersection of top gene hits [3]. The aim of the project was to find the best tool for DE analysis of RNA-seq data of macrophages using Hobotnica. As part of the project, we planned to add additional DE tools into Hobotnica docker container and compare the performance of several DE tools under Hobotnica`s hood.

We complemented original Hobotnica with a bayesian differential expression tool R library baySeq, added features allowing for seamless baySeq volcano-plot and heatmap visualizations along with the already implemented DE analysis tools. Using R language [4] and Rstudio [5] we prepared raw RNA-seq reads GSM4973754 from M0 and M1 macrophages to get thr expression matrix of un-normalised reads counts. We then ran analysis on the prepared data with the new version of Hobotnica.

According to Hobotnica based on the summary H-score on a 0 to 1 scale the 6 DE tools performed as follows: baySeq, NOISeq, EBSeq, edgeR, DESeq, limma-voom scored 0.39, 0.56, 0.58, 0.59, 0.92 and 1, respectively. Thus, it may be assumed that out of 6 DE tools implemented in Hobotnica, limma-voom is the tool of choice in DE analysis of RNA-seq macrophage data. When comparing top 30 genes marked as differentially expressed by each of the tool only one gene (Serine Dehydratase) was intersected between limma-voom, DESeq and NOISeq. The pattern of differential variation revealed by each of the tool was rather variable.

Results of DE tools comparison using Hobotnica expose current problems of existing software in DE analysis. Hobotnica is intended to put some clarity and may be a helpful instrument in the hands of a thoughtful bioinformatician. In the future

Hobotnica may be complemented by more DE tools and features of comparison DE tools on simulated counts data. Validation of Hobotnica results with RT-PCR and proteomics data analysis is upcoming.

References

1. Corchete, L.A., Rojas, E.A., Alonso-López, D. et al. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Sci Rep 10, 19737 (2020). https://doi.org/10.1038/s41598-020-76881-x

2. Costa-Silva J, Domingues D, Lopes FM (2017) RNA-Seq differential expression analysis: An extended review and a software tool. PLoS ONE 12(12): e0190152. https://doi.org/10.1371/journal.pone.0190152

3. Stupnikov A, Sizykh A, Favorov A et al. Hobotnica: exploring molecular signature quality [version 1; peer review: 2 approved with reservations]. F1000Research 2021, 10:1260 (https://doi.org/10.12688/f1000research.74846.1)

4. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

5. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Последняя страница для информации типографии