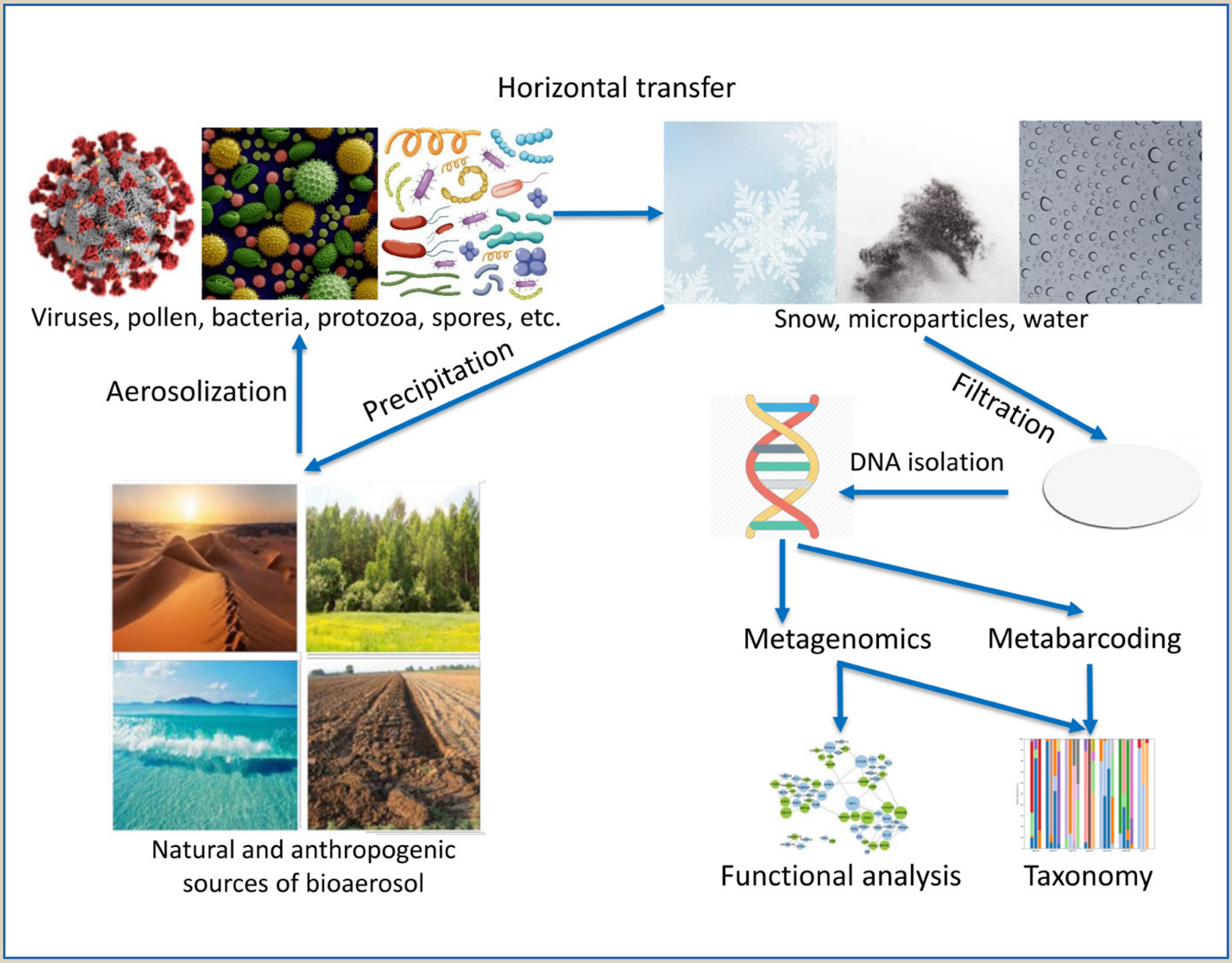
OCTOBER-DECEMBER 2022 VOL. 14 № 4 (55)

ISSN 2075-8251

ActaNaturae

About the Biodiversity of the Air Microbiome


Horizontal transfer
Viruses, pollen, bacteria, protozoa, spores, etc.
Snow, microparticles, water
Aerosolization
Precipitation
Filtration
DNA isolation
Natural and anthropogenic sources of bioaerosol
Metagenomics
Metabarcoding
Functional analysis
Taxonomy

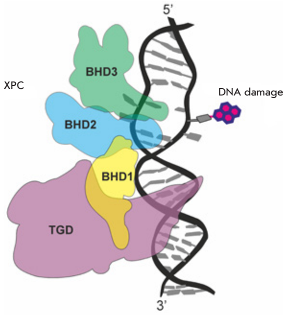
COMPREHENSIVE ANALYSIS OF STROMAL AND SERUM MARKERS IN GASTRIC CANCER
P. 75

MULTIPLE SCLEROSIS IS ASSOCIATED WITH IMMUNOGLOBULIN GERMLINE GENE VARIATION OF TRANSITIONAL B CELLS
P. 84

Bulky Adducts in Clustered DNA Lesions: Causes of Resistance to the NER System

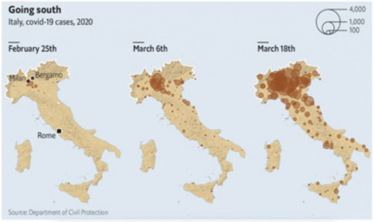
N. V. Naumenko, I. O. Petruseva, O. I. Lavrik

The nucleotide excision repair (NER) system removes a wide range of bulky DNA lesions that cause significant distortions of the regular double helix structure. In this review, we analyzed data on induction of clustered lesions containing bulky adducts, the potential biological significance of these lesions, and methods for quantification of DNA lesions and considered the causes for inhibition of NER-catalyzed excision of clustered bulky lesions.



DNA damage recognition by the XPC protein

The Fallout of Catastrophic Technogenic Emissions of Toxic Gases Can Negatively Affect Covid-19 Clinical Course



Covid-19 map spread in Italy

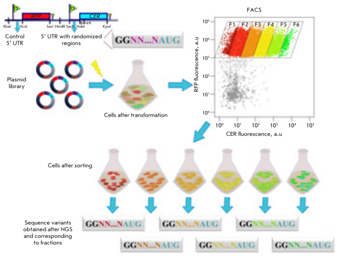
G. Succi, W. Pedrycz, A. P. Bogachuk, A. G. Tormasov, A. A. Belogurov, A. Spallone

The coronavirus D-19 (Covid-19) pandemic has shaken almost every country in the world. In the first wave of the pandemic, Italy suffered an abnormally high death toll. This inexplicably high mortality rate in conditions of a very well-developed health care system such as the one in Lombardy, certainly cries for a convincing explanation. In 1976, the small city of Seveso, Lombardy, experienced a release of dioxin into the atmosphere after a massive technogenic accident. The immediate effects of the industrial disaster did not become apparent until a surge in the number of tumors in the affected population in the subsequent years. In this paper, we endeavor to prove our hypothesis that that release of dioxin was a negative cofactor that contributed to a worsening of the clinical course of COVID-19 in Lombardy.

Flow-Seq Method: Features and Application in Bacterial Translation Studies

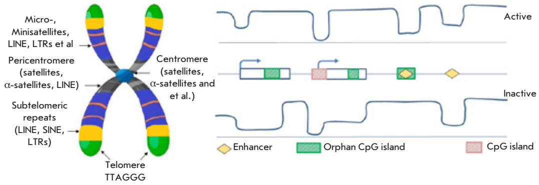
E. S. Komarova, O. A. Dontsova, D. V. Pyshnyi, M. R. Kabilov, P. V. Sergiev

The Flow-seq method is based on using reporter construct libraries, where a certain element regulating the gene expression of fluorescent reporter proteins is represented in many thousands of variants. Reporter construct libraries are introduced into cells, sorted according to their fluorescence level, and then subjected to next-generation sequencing. Therefore, it turns out to be possible to identify patterns that determine the expression efficiency, based on tens and hundreds of thousands of reporter constructs in one experiment. In the presented review, a comparative analysis of the Flow-seq method and other alternative approaches used for translation efficiency evaluation of mRNA was carried out; the features of its application and the results obtained by Flow-seq were also considered.



The scheme of the Flow-seq method

DNA Methylation: Genomewide Distribution, Regulatory Mechanism and Therapy Target


Micro-, Minisatellites, LINE, LTRs et al
Pericentromere (satellites, α-satellites, LINE)
Subtelomeric repeats (LINE, SINE, LTRs)
Telomere TTAGGG
Centromere (satellites, α-satellites and et al.)
Active
Inactive
Enhancer
Orphan CpG island
CpG island

Landscape of the DNA repetitive sequences

D. S. Kaplun, D. N. Kaluzhny, E. B. Prokhortchouk, S. V. Zhenilo

DNA methylation is the most important epigenetic modification involved in the regulation of transcription, imprinting, establishment of X-inactivation, and the formation of a chromatin structure. In this review, the latest research into the DNA methylation landscape in the genome has been summarized to discuss why some DNA regions avoid methylation and what factors can affect its level or interpretation and, therefore, can be considered a therapy target.

ActaNaturae

OCTOBER-DECEMBER 2022 VOL. 14 № 4 (55)
since april 2009, 4 times a year

Founders
Acta Naturae, Ltd,
National Research University
Higher School of Economics

Project Head: N.V. Soboleva
Editor: N.Yu. Deeva
Designer: K.K. Oparin
Art and Layout: K. Shnaider
Copy Chief: Daniel M. Medjo
Web Content Editor: O.B. Semina

Adress: 101000, Moscow, Myasnitskaya Ulitsa, 13, str. 4
Phone/Fax: +7 (495) 727 38 60
E-mail: actanaturae@gmail.com

Reprinting is by permission only.

Номер подписан в печать 30 декабря 2022 г.
Тираж 15 экз. Цена свободная.
Отпечатано в типографии: НИУ ВШЭ,
г. Москва, Измайловское шоссе, 44, стр. 2

Indexed in Pubmed, Web of Science, Scopus, and RISC

Impact Factor: 2.204

CONTENTS

REVIEWS

RESEARCH ARTICLES

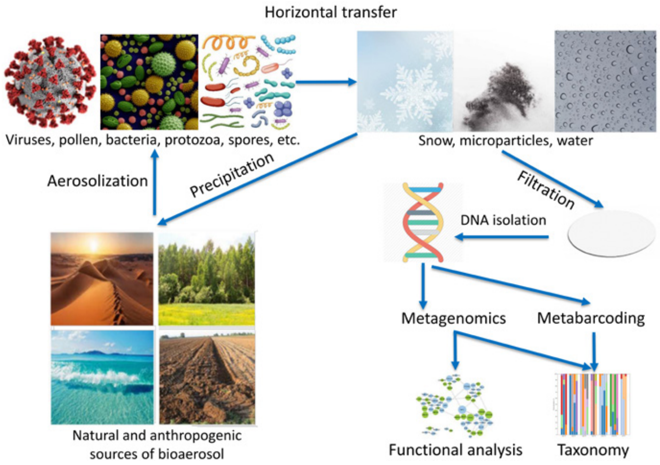

Horizontal transfer
Viruses, pollen, bacteria, protozoa, spores, etc.
Snow, microparticles, water
Aerosolization
Precipitation
Filtration
DNA isolation
Metagenomics
Metabarcoding
Natural and anthropogenic sources of bioaerosol
Functional analysis
Taxonomy

IMAGE ON THE COVER PAGE
(see the article by S. V. Ponomartsev et al.)

DNA Methylation: Genomewide Distribution, Regulatory Mechanism and Therapy Target

D. S. Kaplun[1,2], D. N. Kaluzhny[3], E. B. Prokhortchouk[1,2], S. V. Zhenilo[1,2]
[1]Institute of Bioengineering, Research Center of Biotechnology, Russian Academy of Sciences, Moscow, 119071 Russia
[2]Institute of Gene Biology, Russian Academy of Sciences, Moscow, 119071 Russia
[3]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia
[*]E-mail: zhenilo@biengi.ac.ru

Received October 07, 2022; in final form, November 29, 2022
DOI: 10.32607/actanaturae.11822

ABSTRACT DNA methylation is the most important epigenetic modification involved in the regulation of transcription, imprinting, establishment of X-inactivation, and the formation of a chromatin structure. DNA methylation in the genome is often associated with transcriptional repression and the formation of closed heterochromatin. However, the results of genome-wide studies of the DNA methylation pattern and transcriptional activity of genes have nudged us toward reconsidering this paradigm, since the promoters of many genes remain active despite their methylation. The differences in the DNA methylation distribution in normal and pathological conditions allow us to consider methylation as a diagnostic marker or a therapy target. In this regard, the need to investigate the factors affecting DNA methylation and those involved in its interpretation becomes pressing. Recently, a large number of protein factors have been uncovered, whose ability to bind to DNA depends on their methylation. Many of these proteins act not only as transcriptional activators or repressors, but also affect the level of DNA methylation. These factors are considered potential therapeutic targets for the treatment of diseases resulting from either a change in DNA methylation or a change in the interpretation of its methylation level. In addition to protein factors, a secondary DNA structure can also affect its methylation and can be considered as a therapy target. In this review, the latest research into the DNA methylation landscape in the genome has been summarized to discuss why some DNA regions avoid methylation and what factors can affect its level or interpretation and, therefore, can be considered a therapy target.

KEYWORDS DNA methylation, DNA methyltransferases, G-quadruplexes, TET dioxydenases, methyl-DNA binding proteins.

ABBREVIATIONS TF – transcription factors.

INTRODUCTION

Cytosine is referred to as the fifth DNA base, and cytosine residue methylation is the most common DNA modification in mammalian cells. Cytosine residues in CpG dinucleotides are most often subject to methylation. However, the methylated cytosines outside CpG dinucleotides may account for 25–50% of all mC in stem cells and neurons [1]. In mammals, about 70–80% of cytosines in CpG dinucleotides are methylated [2]. *De novo* DNA methylation is catalyzed by the DNMT3a/3b DNA methylatransferases responsible for methylation in different genome regions and that are not interchangeable [3, 4]. DNA methylation during replication is maintained by DNMT1 DNA methyltransferase. DNA demethylation occurs both passively, during cell division, and actively, due to enzyme activity. The key factors involved in active demethylation are TET1,2,3 dioxygenases. TET proteins oxidize methylcytosine to hydroxymethylcytosine and, then, formylcytosine and carboxycytosine, which then produce cytosine as a result of excision repair by thymine-DNA glycosylase (TDG/NEIL) (*Fig. 1*) [5]. Methylcytosine derivatives are not only considered as intermediate states between methylated and non-methylated bases, but also as DNA modifications affecting the binding of transcription factors, as they are involved in gene expression regulation (Methylcytosine derivatives are discussed in survey [6]).

The key changes in DNA methylation during the organism's development are associated with cell differentiation. Differentiated cells are typically characterized by stable DNA methylation patterns, which can still vary due to external stimuli, various pathological processes, and ageing [7–11]. Dynamic DNA methylation changes in differentiated cells are also observed during memory formation and training in neural cells [12, 13]. DNA methylation in differentiated cells turns out to be stable in the remaining cases. Thus, DNA methylation can be considered as a target for therapy and the diagnostics of various pathogenetic conditions based on DNA methylation abnormalities affecting gene transcription.

The key features of DNA methylation distribution in the genome are presented in this survey. The factors affecting DNA methylation onset, maintenance, and demethylation are analyzed based on recently published data. The possibility for therapeutic use of these factors is discussed.

1. THE DNA METHYLATION DISTRIBUTION PATTERN IN MAMMALIAN CELLS

About 90% of all methylated CpG sistes in mammalian genomes are located in various repeating sequences, such as satellite repeats and mobile elements [14]. The largest number of CpG-rich repeating elements are found in structural chromosomal regions: centromere, pericentromere, and subtelomere (*Fig. 2A*). Genome-wide nonopore sequencing in humans has made it possible to not only read the sequences of repeating elements, but also to analyze their methylation in the genome: so, a significant degree of methylation has been observed under normal conditions [2, 15]. It is of note that methylation of the duplicated/repeating sequences located in various chromosomal regions may differ significantly [2]; i.e., a specific methylation pattern of repeating sequences is not only determined by the sequence repeating it-

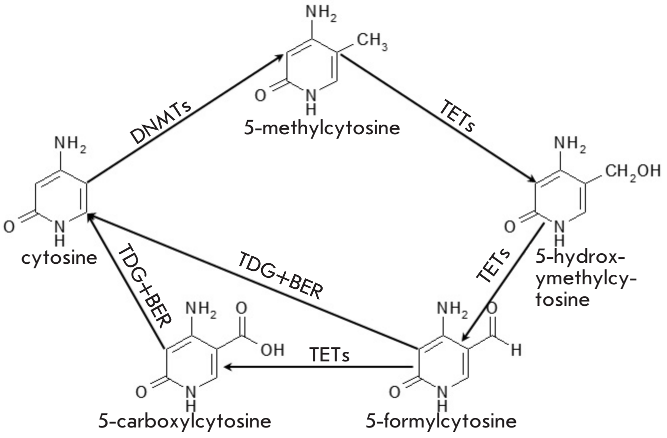


Fig. 1. Cytosine methylation and demethylation scheme

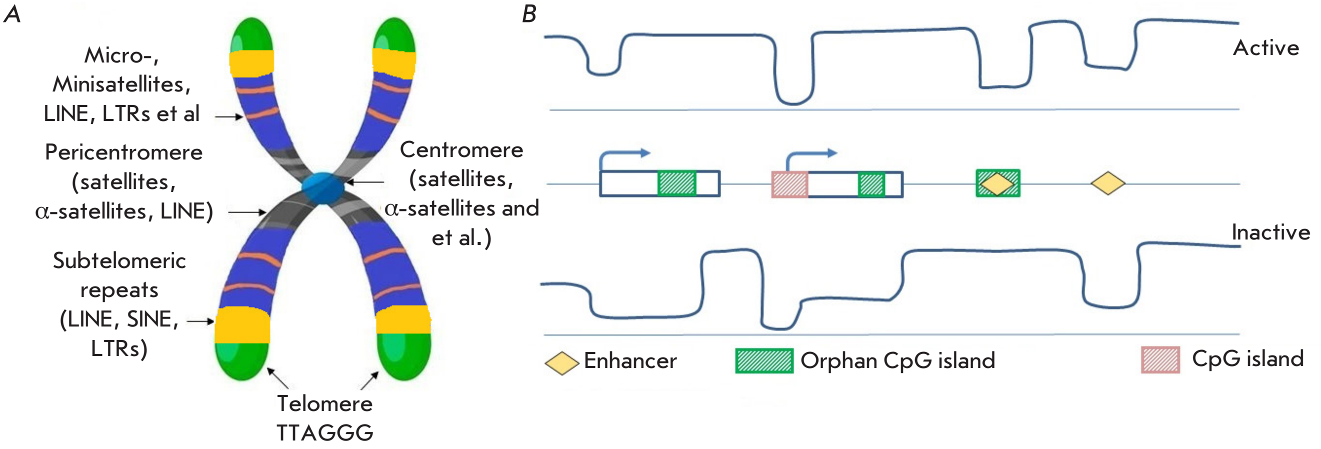

A
Micro-, Minisatellites, LINE, LTRs et al
Pericentromere (satellites, α-satellites, LINE)
Subtelomeric repeats (LINE, SINE, LTRs)
Centromere (satellites, α-satellites and et al.)
Telomere TTAGGG
B
Active
Inactive
Enhancer
Orphan CpG island
CpG island

Fig. 2. Landscape of the DNA repetitive sequences. (*A*) Location of various repeating sequences on a chromosome. (*B*) DNA methylation profile in the genome, depending on the activity of promoters and enhancers and the presence of CpG islands

self, but by its chromosomal surroundings as well. Hypomethylation of various repeating elements is characteristic of various pathological conditions, including oncogenesis, immunodeficiency, as well as autoimmune, neurological, and mental disorders [7–9, 16, 17]. The necessity of satellite repeat methylation in centromere and pericentromere regions is associated with the correctness of chromosome disjunction in replication [18]. In contrast, methylation of mobile elements, transposons, and retrotransposons is aimed at suppressing their transcription. Demethylation of these repeats results in their active transcription and transposition, which fosters genome instability. It is possible that this is a redundant mechanism, because early transposons and retrotransposons are typically characterized by mutations and deletions in the sequences coding for transposase, which leads to inactive protein formation.

The mammalian genome includes CpG dinucleotides that avoid methylation. These CpG sites are usually included in the so-called CpG islands that are DNA regions where the GC pair content exceeds 50%, while the expected-to-observed CpG content is above 0.6. About 60% of the promoters include CpG islands. Lysine 4 residue trimethylation in the histone H3 molecule (H3K4me3) is an active chromatin modification typical for these regions, regardless of promoter activity [19]. Active chromatin is a DNA region where histone modifications, such as acetylation and H3K4me3, lead to DNA accessibility for transcription activators. The presence of H3K4me3 in the promoter regions of inactive genes facilitates transcription initiation but not mRNA synthesis. Meanwhile, there is a series of inactive gene promoters, including non-methylated CpG islands, where H3K4me3 is not detected. The genes located in clusters with three or more homologous genes coding for olfactory receptors, keratins, apolipoproteins, interleukins, and leukocyte antigens are most commonly attributed to this class [19]. Methylation of CpG islands in the promoter regions correlates with transcription suppression and may occur both under normal (e.g., during organism development) and pathological conditions [20]. For instance, malignant cell transformation and metastasis are typically characterized by hypermethylation of CpG islands in the promoters of oncosuppresor genes; i.e., the proteins involved in cellular adhesion and DNA repair. In most cases, such hypermethylation results in transcription suppression. However, it should be noted that promoter hypermethylation in tumors may occur in the genes considered transcriptionally inactive in the same tissue under normal conditions. In other words, their hypermethylation has no effect on expression suppression but, rather, reinforces their inactive status [21].

Promoters that include a small quantity of CpG dinucleotides are typical of tissue-specific genes and the genes involved in organism development. Methylation in these promoters does not always correlate with transcription suppression [22]. Inn a comparative analysis of brain and retinal cells, methylation of 66% of differentially methylated promoters correlated negatively with transcription. Thus, methylation in these promoters corresponds to transcription suppression. At the same time, promoter methylation was observed in 34% of transcriptionally active genes in [22].

The CpG islands that do not overlap with promoter regions are called orphan islands. The number of such islands is about half that of the promoter islands. Orphan islands often include H3K4me3 active chromatin modification and can initiate new transcripts [23]. Many orphan islands are subject to methylation during organism development, which makes them lose active chromatin modifications. Methylation in an orphan island inside a gene prevents the occurrence of transcription initiation sites inside the gene and correlates with active transcription [24]. Methylation inside genes may prevent Polycomb protein binding in the PRC2 repressor complex, which facilitates active transcription as well [25]. About 90% of orphan islands may act as tissue-specific enhancers [26]. The presence of a CpG island amplifies the enhancer's regulatory activity [27]. Active enhancers that include orphan islands are hypomethylated, while the classical enhancers operating in all tissue types show variable methylation [27] (*Fig. 2B*).

Methylation maps created for the whole genomic DNA in various cell types and the information regulatory activity of the elements make it possible to consider DNA methylation as a tool for transcription activity regulation with regard to correction or identification of the various pathogenetic states associated with changes in DNA methylation.

2. DNA METHYLATION HOMEOSTASIS

DNA methylation homeostasis is based on a complex regulatory network that balances methylation and demethylation. The key mechanisms maintaining homeostasis in cellular proliferation and differentiation are as follows: 1) passive genome-wide demethylation and maintainance of the methylation pattern by DNMT1 during replication and 2) targeted *de novo* methylation and active demethylation in specific regions. The factors involved in homeostasis are discussed in the present Chapter.

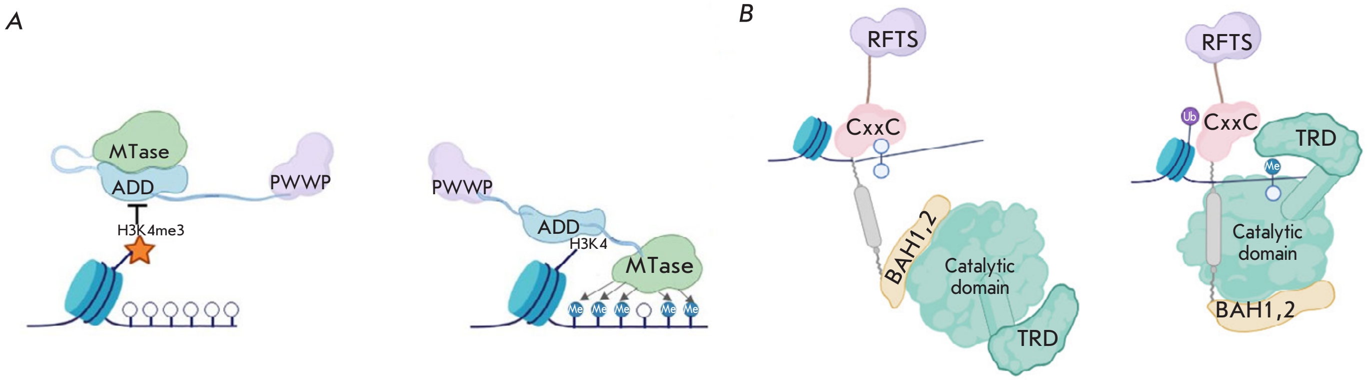


Fig. 3. Binding scheme (*A*) DNMT3a/3b to unmodified H3K4, the presence of H3K4me3 prevents the ADD domain from binding to DNA, which leads to autoinhibition of the enzyme; (*B*) DNMT1 to DNA, the interaction with unmethylated DNA leads to inhibition of the catalytic domain [29, 30]

2.1. Maintaining the non-methylated state in DNA regions

About 20% of CpG dinucleotides, most of them CpG islands, avoid methylation. The main factors preventing their methylation include histone modifications, DNA intercations with certain transcription factors (TF), and the DNA primary and secondary structure.

2.1.1. H3 lysine 4 trimethylation. The presence of trimethylated H3K4 is among the reasons explaining the stability of CpG islands against *de novo* methylation regardless of the transcriptional activity of the region. H3K4me3 prevents the attraction of *de novo* DNA methyltransferases DNMT3a/3b and their regulator, DNMT3L, showing no own catalytic activity to the DNA [28]. DNA methyltransferases DNMT3a/3b include a catalytic domain showing methyltransferase activity (MTase), as well as ADD and PWWP domains involved in chromatin binding. DNA methyltransferases, when in their DNA-unbound form, are inactive due to autoinhibition: the ADD domain interacts with the catalytic domain hindering its activity (*Fig. 3*). The ADD domain is unable to interact with H3K4me3. At the same time, non-modified H3K4 interacts with the ADD domain of DNMT3a/3b, thereby disrupting ADD binding to the catalytic domain and facilitating the manifestation of methyltransferase activity [28, 29]: Thus, DNA methylation and H3K4 methylation are mutually exclusive phenomena (*Fig. 3A*).

The case when DNA methylatransferase DNMT1 is maintained is different. DNMT1 is localized in promoter regions, including non-methylated CpG islands, and is not involved in their methylation. DNMT1 includes the following domains: RFTS (replication foci-targeting sequence), ZF-CxxC, two BAH- (bromo-adjacent homology) domains, and the catalytic domain. The CxxC domain of DNMT1 may bind to sequences including non-methylated CpG dinucleotides. Meanwhile, the BAH1 domain physically intervenes through the interaction between the catalytic domain and the DNA, thereby preventing *de novo* methylation (*Fig. 3B*) [30].

A particular feature of CpG islands is their ability to bind to TF and enzymes containing the ZF-CxxC domain (CFP1, MLL1/2, KDM2A/2B, TET1/TET3, DNMT1) [31]. Many of these factors are represented by or bind to the histone methyltransferases that modify H3K4, which hinders the attraction of DNA methyltransferases. It should be noted that the lower the gene promoter activity, the higher the need for H3K4me3 to maintain its non-methylated state [32, 33].

2.1.2. TET dioxygenases. TET dioxygenases (ten-eleven translocation) are the enzymes that oxidize methylcytosine for the subsequent excision repair. TET proteins are attracted to the DNA through various mechanisms. TET bind to the CpG islands by their CxxC domain or other transcription factors with a CxxC domain. TET proteins may also be attracted to DNA without the involvement of CpG islands, through messenger proteins, such as Klf4, Nanog, REST, GADD45, CEBPa, etc.; e.g., TET1 and TET2 are attracted to DNA by binding to the TF Nanog, leading to the demethylation of the regulatory gene regions involved in the maintenance of the pluripotent cellular state [34]. Notably, TET proteins, similar to many CxxC-containing proteins, affect H3K4 trimethylation. TET interact with OGT transferase (O-GLCNac transferase), which in turn forms a complex with SET1 and MLL histone methyltransferases trimethylating H3K4 [35].

The so-called pioneer factors play a major part in DNA demethylation by TET dioxygenases [36]. They interact with closed, inactive chromatin and change its accessibility for transcription activators. They show their peak activity during organism development, immune system maturation, oncogenesis, and somatic cell reprogramming. Pioneer factors include FOXA1, FOXO, Sox, Pax, GATA, Oct4, PU1, CEBPα, and other TF [37]. The key feature of these factors is their ability to recognize not just a DNA sequence, but a DNA region in a nucleosome context as well [38, 39]. This explains why DNA methylation is not always critical for pioneer factor attraction. In fact, many pioneer factors are methylation-insensitive or have recognition sites that do not contain CpG dinucleotides, which is illustrated by the cases of ASCL1 and FOXA1 [40, 41]. Nevertheless, the pioneer factors Oct4 and Klf4 interact both with sequences not containing CpG and sites containing CpG. In the latter case, Oct4 and Klf4 only bind to the methylated sites [42]. The pioneer factors capable of forming complexes with TET dioxygenases include Klf4, CEBPa, and TFCP2l1 [37]. The functional significance of TET2 interaction with Klf4 and CEBP in the process of somatic cell reprogramming has been demonstrated: e.g., the pioneer factors Klf4 and CEBPa attract TET2 dioxygenase to methylated enhancer sequences, which leads to their demethylation and activation [37]. Here, methylation decrease in certain chromatin regions, including in the Klf4 binding sites, is followed by chromatin remodeling. *TET2* knockout cells are not subject to reprogramming [37]. Thus, DNA demethylation by TET enzymes is among the key stages of cellular reprogramming.

Despite the involvement of TET proteins in demethylation in many regions, their removal does not lead to catastrophic changes in the genome-wide DNA methylation level. The main DNA methylation changes in the case of TET knockout have to do with distal regulatory elements and enhancer sequences [43].

2.1.3. DNA secondary structure. Changes in conformation – aka DNA secondary structure – are among the factors contributing to the maintenance of a non-methylated state in CpG islands.

One of these factors is an R-loop, which is an RNA-DNA hybrid and a displaced DNA strand. GADD45A binding to an R-loop in the promoter of the tumor suppressor gene TCF21 attracts TET1, facilitating local demethylation in the region [44]. Thus, the DNA secondary structure may affect DNA demethylation by binding to TET dioxygenases.

G-quadruplexes can also have an effect on the methylation of CpG islands and the CpG dinucleotides not included in the islands. It is an established fact that the GC-rich regulatory regions of eukaryotic genomes are capable of changing local DNA conformation by arranging themselves into alternative structures in the form of G-quadruplexes (G4) [45]. The secondary G-quadruplex (G4) structure is formed by guanine-rich sequences. The G-G base-paired Hoogsteen interaction results in guanine quartet formation, and stacks of such quartets stabilized by potassium cations form the G4 core. The thermodynamic stability of these structures depends on the nucleotide sequence and sometimes exceeds that of the DNA double helix. There are several theoretical and experimental approaches to determining potential G4 regions. Stable G4s formed in the genomic DNA in the presence of potassium ions act as the barrier for DNA polymerase. It often becomes an obstacle for PCR amplification in genome regions including GC-rich sites prone to the formation of G4 structures [46]. The approach based on high-performance sequencing of the errors occurring in the presence of potassium ions is currently considered the best in the experimental prediction of the G4 refolding potential in genomic DNA [47]. A change in the DNA conformation affects its physical and chemical properties and the affinity of various proteins specific to a certain nucleotide sequence. Methylation in the CpG context may change the energy barrier for a transition between the DNA double helix and non-canonical DNA structures, in particular G4 [48]. About 30% of CpG islands include nucleotide sequences capable of forming G4 structures (*Fig. 4A*). Intragenic CpG islands are relatively rich in quadruplex sequences, while the probability of their occurrence in intergenic CpG islands is low. The highest G4 density significantly above the average for all promoters (*Fig. 4C*) is detected in promoter CpG islands (*Fig. 4B*). The maximum G4 density is observed near the transcription start site (TSS). Decreasing G4 occurrence in promoter regions with no CpG islands may be related to the differences in the GC-contents between the promoters overlapping with CpG islands and those removed from them (*Fig. 4D*). The probability of encountering a potential G4-quadruplex

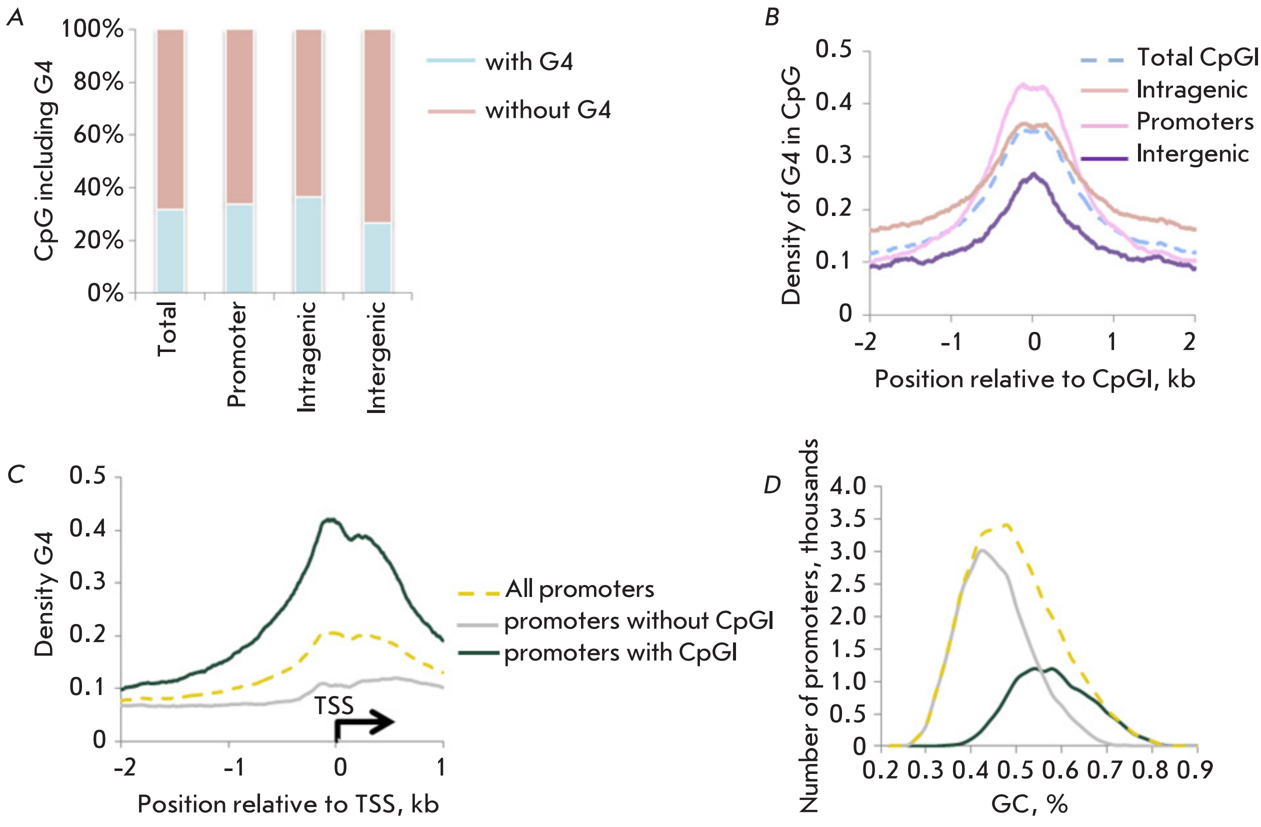


Fig. 4. Distribution of potential G4 sequences in CpG islands. (*A*) The proportion of CpG islands with G4, (*B*) the distribution density of G4 near CpG islands depending on the localization in the genome. (*C*) G4 density and (*D*) GC composition in promoter regions depending on the presence of CpG islands

depends significantly on the GC-content, even in a randomly generated nucleotide sequence. The probability of encountering a potential G4 in a random sequence with a GC-content of 40% is about one G4 per a million base pairs, while increasing the GC-content in a random sequence to 70% increases the G4 occurrence probability to one per one thousand base pairs [49]. The probability of encountering a G4 sequence in a higher organism genome is above average. These sequences may have an important regulatory role, which is somewhat confirmed by positive evolutionary selection [50]. The presence of G4 in promoters is often associated with transcription suppression [51]. Nevertheless, G4 in stem cells is detected in active promoters and the sites interacting with them; i.e., enhancers, superenhancers, and TF binding sites determining the cell type. In addition to active regulatory elements, G4s are found in regions with bivalent chromatin modifications; i.e., the ones containing both active and inactive chromatin modifications. A decrease in the detected G4 structures associated with cell differentiation correlates with the occurrence of a closed chromatin [48, 52]. Quadruplex structures may interact with DNA-methyltransferases DNMT1, DNMT3A, and DNMT3B *in vitro* [53, 54]. Indeed, non-methylated sequences in CpG islands containing quadruplexes are rich in DNMT1 binding sites. Meanwhile, interaction between DNMT1 and G4 leads to its DNA methyltransferase inactivation [53]. Thus, G4 formation hinders DNA methylation. This is confirmed by the correlation between the presence of stable quadruplexes in open chromatin and DNA hypomethylation. This correlation is primarily characteristic of sites with a low GC content. Relatively low methylation is also typical for CpG islands in a closed chromatin containing quadruplexes, compared to regions free of quadruplexes [55].

It is still unclear what DNMT1 activity – specifically the binding to non-methylated CpG sites, when domain positioning hinders the catalytic activity, or interactions with non-canonical DNA structures – is critical in maintaining the non-methylated status of CpG islands. Notably, there are genome regions where DNMT1 binding to DNA manifests *de novo* methyltransferase activity. These regions include LTR-retrotransposons enriched with H3K9me3 and TRIM28. Here, DNMT1 *de novo* activity is regulated by UHRF1 [56]. Thus, the presence of co-factors is also critical for the manifestation of DNMT1 *de novo* activity, in addition to domain positioning.

2.1.4. Competition between transcription factors and DNA methyltransferases. TF binding to DNA can occur with attraction of DNA methylatransferases, thereby protecting DNA from methylation. Sp1 is the classic example of this competition between TF and DNA methyltransferase binding. Sp1 interacts with the non-methylated sequences CCGCCC CpG islands are enriched with and intervenes, with attraction of DNA methyltransferase [57]. Mutation in an Sp1 binding site leads to its increased methylation and reduced transcription [58]. Thus, Sp1 is considered as TF obstructing the methylation of CpG islands. However, unavailability of a recent genome-wide analysis of DNA methylation with Sp1 removed makes it impossible to confirm whether Sp1 is necessary for maintaining the non-methylated status in multiple CpG islands.

CTCF is another factor contributing to the maintenance of a non-methylated DNA state. CTCF is identified as TF binding to non-methylated sequences and capable of acting both as transcription activator and repressor. CTCF also acts as insulator; i.e., it blocks enhancer action on promoters and, therefore, is involved in chromatin structure formation [59]. CTCF binds to non-methylated alleles in imprinted loci, disrupting the enhancer-promoter interaction. CTCF binding to the non-methylated maternal allele in the H19/Igf2 locus is not only critical in terms of enhancer-promoter interaction, it also affects the maintenance of the maternal allele in a non-methylated state. Mutations in the CTCF binding sites in this locus resulted in increased methylation of the maternal allele after ovum fertilization, but methylation in the H19/Igf2 locus in germ cells was not disrupted in [60]. CTCF reduction in the oocytes mediated by RNA interference (RNAi) resulted in increased maternal allele methylation in the locus of interest in [61, 62]. Thus, CTCF turns out to be critical in maintaining the maternal allele in the H19/Igf2 locus in a non-methylated state. CTCF loss in cancer cells leads to hypermathylation in the protein binding sites as well [63]. According to the genome-wide analysis, CTCF is primarily localized in the non-methylated or poorly methylated regions in the stem cells of mice. Nevertheless, some CTCF binding sites are found to be highly methylated [64]. It turns out that methylation intervenes with the CTCF-DNA interaction only at specific positions of the binding site [65]. Mutation in the methylated CTCF binding sites does not cause changes in the methylation level, despite the fact that the presence of CTCF in methylated sequences correlates with lower methylation compared to the regions lacking CTCF recognition sites [66]. Thus, the interaction between CTCF and methylated sequences has nothing to do with maintainance of the methylation level in these regions. It should be noted that DNA methyltransferase knockout cells with a reduced DNA methylation level showed no redistribution of CTCF binding sites onto demethylated regions in [67]. Thus, the DNA methylation, on its own, is not an obstacle to CTCF binding. The sites were found in the imprinted H19/Igf2 locus, which CTCF can bind to *in vitro* regardless of their methylation level. It is possible that CTCF is not detected on the methylated allele *in vivo* due to competing binding of methyl-sensitive proteins [68]. Thus, CTCF shows varying DNA binding activity but binding to non-methylated sequences maintains the sequences' low methylation level.

The search for factors protecting DNA from hypermethylation, akin to CTCF or Sp1, could make it possible to study new mechanisms for maintaining the DNA in a non-methylated state and consider them as targets for manipulating DNA methylation and the transcription activity of genes in conditions associated with DNA methylation abnormalities.

2.2. Maintaining DNA regions in methylated state

In this chapter, the processes of DNA methylation onset and maintenance are discussed. They are critical to various repeating sequences, imprinted sites, and regulatory elements. *De novo* DNA methylation involves the DNMT3a and DNMT3b methyltransferases, but, as mentioned above, DNMT1 may manifest *de novo* activity as well. DNMT3a DNA methyltransferase is responsible for methylation onset in the repeating sequences, regulatory elements, and gene bodies acting as Polycomb protein targets. DNMT3b is critical for methylation onset in the regions of satellite repeats and sequences on inactivated X chromosomes [3, 4]. Histone modifications and interactions with transcription factors are important for DNA methytransferase attraction. Long non-coding RNA and PIWI-interacting non-coding RNA can act as messengers regulating *de novo* methyltransferase binding to DNA as well [69].

2.2.1. Histone modifications. DNMT3 attraction to DNA is achieved using various mechanisms, including histone modifications. As mentioned above, non-modified H3K4 facilitates the binding of DNA methyltransferases through the ADD domain and amplifies their catalytic activation. In addition, DNA methylation is regulated by H3K36me3/me2 histone modifications. DNMT3 methylates the CpG-rich intragenic sequences of actively transcribed genes in the regions characterized by the presence of H3K36me3-modified histones. DNA binding and DNMT3a-mediated methylation in intergenic regions requires H3K36me2. The PWWP domain of DNMT3 is responsible for the interactions with H3K36me2/me3 [70, 71].

DNMT3 binding to heterochromatin and repeating sequences is mediated by H3K9 methylation. DNA methyltransferases are attracted to DNA due to interaction with the histone methyltransferases methylating H3K9 (Suv39h1/2, G9a/GLP, Setdb1) and binding to the HP1α and HP1β proteins recognizing methylated H3K9 [72].

2.2.2. Transcription factors attract DNMT to DNA. DNMT3a and DNMT3b are not interchangeable, and their mutations and deletions result in methylation changes in general and specific regions [3, 4]. This has to do with the fact that they are attracted to DNA through interaction with various TFs. As of now, a lot of TFs are being discovered which are capable of interacting with one or both DNA methyltransferases or can be included in a complex with them without interacting directly [73]. Interestingly, these TFs only affect methylation in a limited number of direct targets, which are in many cases restricted to individual target genes. As a result, these TFs may be considered as targets for selective regulation of target gene methylation. Let us discuss some of these factors.

GCNF

GCNF (germ cell nuclear factor) participates in methylation onset and maintenance in various promoter regions by directly interacting with DNMT3a/3b methyltransferases [74]. In addition, GCNF may indirectly attract DNMT3 methyltransferases. GCNF in stem cell differentiation binds to the *Oct4* promoter and interacts with MBD2 and MBD3, which in turn are included in a single complex with DNMT3. This leads to *Oct4* methylation and its transcription suppression in differentiated cells. Since MBD2/MBD3 cannot bind to *Oct4* during stem cell differentiation with GCNF knockout, the gene remains active [75]. GCNF ability to regulate *Oct4* methylation may be used to analyze a cellular pluripotency status. For instance, *GCNF* promoter demythilation is observed in somatic cell reprogramming, which enables gene activation during cell differentiation that effectively suppresses *Oct4* transcription. These pluripotent cells are mature, but if their reprogramming is not completed, then *GCNF* promoter methylation is maintained, gene activation does not occur during cell differentiation, and *Oct4* remains active in differentiated cells, rendering them potentially oncogenic. Thus, GCNF, or more specifically its promoter methylation, can be considered a maturity marker for pluripotent cells.

Kaiso (ZBTB33)

Proteins containing a zinc finger domain often act not only as methyl-DNA-binding proteins, but also as factors contributing to DNA methylation homeostasis [42, 76]. A particular feature of these proteins is their ability to recognize both methylated and non-methylated regions often different in terms of their nucleotide sequences. The zinc finger structure makes it possible to specifically recognize a methylated CG site, most often in a certain context for each TF [77]. The first established proteins to include zinc finger domains interacting with methylated sequences were Kaiso-like proteins: Kaiso (ZBTB33), ZBTB4, and ZBTB38. In addition to zinc fingers, they include the BTB/POZ domain responsible for the protein-protein interaction at their N-end [78–80]. Later, other zinc-finger proteins capable of interacting with the methylated DNA were discovered, including Znf57, CTCF, Klf4, Wt1, and Egr1. The strongest affinity to the methylated DNA is demonstrated by Kaiso and Znf57, binding to methylated sequences over 20 times better than to non-methylated sequences. At the same time, the sensitivity to methylated sequences in the remaining zinc-finger proteins is only 1.5-3 times as high or equal to that for non-methylated sequences [81, 82].

Kaiso binds to methylated sequences and regions including CTGCNA [78, 80]. This protein can act as a transcription repressor, with the BTB/POZ domain at the N-end attracting the NcoR and SMRT corepressor complexes, and as transcription activator [83–85]. Imprinted H19/Igf2 locus is a target for Kaiso that binds to the methylated allele of the locus, and its removal results in ICR1 methylation decrease in the locus [86, 87]. It is possible that methylation decrease following Kaiso removal is due to competition with CTCF, which in turn can bind to methylated sequences and cause their demethylation. In case of Kaiso knockout, methylation decrease is observed in the *Oct4* promoter in the embryonic fibroblasts of mice and the *TRIM25* promoter in human embryonic renal cells, gene bodies, enhancers, and regions not containing histone modifications [83, 88, 89]. It is shown that *TRIM25* promoter demethylation caused by Kaiso removal is reversible by the expression of exogenous Kaiso, which can be included in a complex with DNMT3a/3b [83, 89]. Notably, Kaiso removal in cancer renal cells in humans causes a slight genome-wide methylation increase. This uniform distribution may be associated with the decrease in TET1 dioxygenase transcription; i.e., Kaiso can shift DNA methylation in both directions. Thus, Kaiso not only maintains the required methylation level, but also participates in methylation onset in various loci by interacting with DNA methyltransferases 3a and 3b [89].

The regulating role of Kaiso in DNA methylation may also be associated with its ability to interact with the ubiquitin-like proteins SUMO1,2,3. The SUMO proteins covalently bind to lysine residues in the target proteins, similarly to ubiquitin. Unlike ubiquitination, SUMOylation usually does not cause protein degradation, while affecting cellular localization, activity, and interaction with other factors. Kaiso SUMOylation affects its transcription properties [83]. The presence of six SIM-SUMO interacting motifs in the Kaiso amino acid sequence and non-covalent interaction between Kaiso and SUMO1 allow us to assume that Kaiso can act as a E3 SUMO ligase. SIM sites are sequences of several hydrophobic amino acid residues surrounded by serine or acidic amino acid residues. The so-called non-canonical E3 SUMO ligases include SIM and non-covalently interact with SUMO [90]. Many proteins are SUMOylated in so-called PML and/or PcG bodies [90, 91]. Kaiso is localized in PcG bodies in the case of exogenous SUMO expression [92]. This allows us to assume that Kaiso not only participates in transcription regulation and DNA methylation maintenance, but may also participate in activity regulation of other factors by affecting their post-translation modifications. For example, SUMOylation of DNA methyltransferases increases their catalytic activity, thereby facilitating an increase in DNA methylation [93]. On the other hand, SUMOylation of the XRC11 excision repair protein is required for effective removal of 5-formyl- and 5-carboxycytosines in stem cell differentiation and subsequently effective DNA demethylation [94]. That is why studying Kaiso in terms of E3 SUMO ligase and searching for its potential targets makes it possible to uncover new activity regulation mechanisms for various factors, including the proteins contributing to DNA methylation.

Znf57

Unlike Kaiso, Znf57 contains a KRAB (Krueppel-associated box) domain at the N-end. Znf57 binding to methylated sequences using the KRAB domain attracts the TRIM28 (KAP1) corepressor, which forms a complex with H3K9 histone methyltransferase SETDB1 and DNA methyltransferases DNMT1 (maintenance) and DNMT3a/3b (*de novo*) [95]. This repressor complex is formed in the transposon region, imprinted loci, and on inactive enhancers [96, 97]. Znf57 removal causes demethylation in imprinted loci and embryonic death [96]. It should be noted that Znf57 is responsible for methylation maintenance, but not onset.

UHRF1

UHRF1 plays a key role in DNA methylation maintenance in replication. This explains its expression pattern: UHRF1 is only detected in actively dividing cells (for example, spinal cord cells), where DNA methylation onset in the daughter strand is required in replication, and not detected in terminally differentiated cells (neurons, hepatocytes). UHRF1 binds to methylated and semi-methylated DNA using the SRA domain (SET and RING- associated domain). UHRF1 also includes several domains participating in protein-protein interactions: UBL (ubiquitin-like domain), TTD (tandem tudor domain), PHD (plant homeodomain), and RING (a really interesting new gene domain). These domains ensure interaction with maintenance DNA methyltransferase DNMT1, PCNA, histone deacetylase HDAC1, histone methyltransferases G9a, and SUV39H1, PARP1, etc. [98]. UHRF1 binding to semi-methylated DNA in replication ubiquitinates H3K18 and H3K23 and attracts DNMT1 methyltransferase for methylation establishment in the daughter DNA strand. DNMT1 activity is regulated by interaction with H3K18ub and H3K23ub [99]. In pathogenetic tumor conditions, UHRF1 may also affect methylation onset in the promoters of some genes [100]. UHRF1 removal leads to genome instability, G2/M phase arrest, and apoptosis. The absence of double strand break repairs is observed in cells as well [101]. Thus, UHRF1 contributes to DNA methylation establishment and maintenance.

MBD proteins

Methyl-DNA binding proteins with MBD (methyl DNA binding domain) are found among proteins not only recognizing methylated DNA, but also contributing to the binding site methylation. Most MBD proteins are involved in the formation and functioning of the nervous system. There are only four factors in this family (MBD1, MBD2, MBD4, and MeCP2) capable of binding to methylated DNA. These MBD proteins show the strongest affinity to methylated CpG islands [102]. In most cases, these proteins act as interpreters of methylation: i.e. they attract corepressors or compete with transcription activators for DNA binding. However, some recent studies show that these factors can also contribute to DNA methylation establishment and maintenance. It was shown that MeCP2 knockout leads to the occurrence of both hypo- and hypermethylated regions in various types of neurons in mice [103]. The mechanism behind the effect of MBD proteins on the methylation level is yet to be studied. MBD1 regulates methylation in the promoters of the Htr2c serotonin receptor gene and bFGF growth factor [104, 105]. MBD1 knockout leads to Htr2c reactivation, which is considered among the causes of deviations in hippocampal neurogenesis, learning disorders, and occurrence of autism symptoms associated with social behavioral changes, attention deficit, and serotonin activation abnormalities in gene-knockout animals [104]. Reactivation of the bFGF growth factor in the case of MBD1 knockout affects the ability to maintain the pluripotent state of stem cells, whose regulation is important for the subsequent differentiation into cells of the nervous system [105].

Thus, there are factors, such as UHRF1, that contribute to genome-wide methylation maintenance and ones (MBD, Kaiso, and GCNF proteins) that regulate methylation in a specific variety of targets. The latter are of special interest, since identification of their binding sites, whose methylation level is affected by the inactivation or mutations of these factors, could make it possible to manipulate the methylation levels in their targets by changing their activity. Interestingly, the desired changes in DNA methylation may also be regulated by activity modulation of DNA methyltransferase using post-translation modifications: Kaiso is a potential E3 SUMO ligase.

3. DNA METHYLATION EDITING

Using advanced DNA editing methods to change methylation levels in certain regions is one of the ways used to alter their transcription activity. This approach makes it possible to alter promoter and enhancer activity using mutated dCas9 endonuclease incapable of DNA cutting. To ensure DNA hypermethylation, dCas9 is bound to the catalytic domain of DNMT3, whose methyltransferase activity is targeted on the region of interest [106]. The TALEN and zinc finger domain may be used instead of dCas9, but the editing system based around dCas9 remains the most accessible one. The main problems with this editing technique are as follows: 1) the methylation level is not high enough, and 2) DNA demethylation occurs after a certain number of cell divisions. To solve these problems, DNMT3L acting as a cofactor amplifying DNA methylation is added to the catalytic domain of DNMT3. A high methylation level is maintained during a lasting cell division by introducing the chimeric construct dCas9-Ezh2 or dCas9-KRAB into the cells. Ezh2 trimethylates H3K27, and the KRAB domain of Znf57 acts as a base that can be used to assemble a repressor complex that modifies histones and methylates DNA [107]. It is also necessary to identify which factor – Ezh2 or KRAB – would be more effective in suppressing the transcription activity in the region of interest [107].

To ensure DNA demethylation, dCas9 is bound to the catalytic domain of TET proteins [108]. Introduction of the catalytic domain of TET causes not only demethylation, but also 5-hydroxymethylcytosine formation, which contributes to TF attraction [109]. To achive a more reliable DNA demethylation impact (without cytosine intermediates), dCas9 may be bound to the catalytic domain of ROS1 DNA glycosylase *Arabidopsis* [110].

The key advantage of DNA methylation editing, compared to DNA editing, is that the nucleotide sequence remains intact while only DNA modification changes. These changes are reversible, and almost any sequence in the genome can be edited.

4. DNA METHYLATION AND PATHOGENETIC CONDITIONS

In recent years, the relationship between the regulatory mechanism of DNA methylation and various pathogenetic conditions, especially oncogenesis, rheumatoid arthritis development, and various neurological diseases, has been uncovered [11, 111]. Two categories of clinical significance of the changes in the DNA methylation level can be identified. The first one includes the cases where DNA methylation may act as a marker for a developing pathogenetic condition. The second one includes cases where changes in DNA methylation and the activity of methyl-DNA binding proteins affect the course and progression of the condition.

4.1. DNA methylation as a diagnostic and predictive marker of disease progression

The DNA regions whose methylation changes can be detected in damaged organs or tissues, blood genomic DNA, DNA from various body fluids, and circulating-free DNA are selected as markers of disease progression. Markers making it possible to quite accurately predict oncological diseases at their early stages, evaluate the effect of therapy, detect recurrent cases, and even identify tumor types in some cases, have been selected [112–114].

4.2. DNA methylation as a target for therapy in various pathogenetic conditions

Hypermethylation in the CpG islands located in suppressor gene promoters, leading to their inactivation, is often detected during oncogenesis. Tumor suppressor genes can be activated, albeit inconsistently, through promoter demethylation. For instance, 5-azacitidine reducing DNA methylation is used as an active substance in decitabine used as therapy in acute myeloid leukemia and myelodysplastic syndrome. However, instead of targeting a specific gene, this drug affects the whole genome, causing its instability and damaging the DNA, which may have severe consequences for the patient [115]. Methylation in the promoters of tumor suppressor genes may be reduced by inactivation of the catalytic activity of the maintenance DNA methyltransferase. Inhibitors of DNMT1 DNA methyltransferase RG108 and SG102 are less toxic than 5-azacitidine. They do not change methylation in satellite repeats but affect promoter demethylation, including in some suppressor genes [116, 117]. The key limitation of these inhibitors is the small quantity of targets; i.e., the regulatory elements of suppressor genes. The catalytic activity of DNMT1 may also be suppressed using oligonucleotides that form quadruplex structures [53]. Attempts are made to manipulate DNA methylation using the editing system. The bottleneck of this approach is the delivery of dCas9 or its analogues to target organs and tissues [118]. Hepatocytes, where the editing system can be delivered via injection (for example, tail vein injection in mice), are one of the most accessible targets. Attempts to reduce methylation in the *Fgf21* promoter in the liver of mice have been described. *Fgf21* codes for the factor participating in glucose and cholesterol metabolism. Introduction of dCas9 with the catalytic domain of TET1 resulted in a short-term methylation decrease in the promoter on the sixth day after injection, and as early as the 14th day the methylation level was restored in [119]. Thus, stable DNA methylation editing in a living organism is yet to be achieved.

4.3. Methyl-DNA binding proteins as new targets for therapy

When selecting a therapy target, one should take into consideration how critical is the inactivation of a factor to the organism. Knockout or mutations in the methyl-DNA binding proteins MBD1, MBD2, MeCP2, and Kaiso result primarily in behavioral deviations not disrupting vital processes, which may be reversed upon restoration of protein expression as in the case of MeCP2 [120, 121]. Inactivation of these proteins changes the general methylation level insignificantly and does not lead to genome instability and reactivation of repeating elements. Hence, the MBD proteins Kaiso and their homologue ZBTB4 enjoy an advantage as potential targets. The search for the target genes of these factors associated with pathogenetic conditions seems a promising line of research.

For instance, investigation of the binding sites in methyl-DNA binding proteins made it possible to identify the gamma globin gene as a methyl-dependent target. A gradual transition of hemoglobin types occurs during the human organism's development: the epsilon globin gene is transcribed in the embryonic period; gamma globin – at birth; and beta globin – in adulthood. Patients with the sickle-cell disease and beta thalassemia show an abnormal expression of or mutations in the beta globin gene, leading to severe consequences. Reactivation of the normal form of gamma globin would make it possible to restore a normal hemoglobin level in the blood. The methyl-DNA binding protein MBD2 regulates the attraction of the NuRD corepressor complex to the promoter of the gamma globin gene in blood cells and maintains it in an inactive state in adults [122]. MBD2 removal leads to a 20-fold increase in the expression of the gamma globin gene [123]. Transcription of the gamma globin gene may be activated by disrupting MBD2 binding to the NuRD corepressor complex and its components using inhibitors (*Fig. 5*). Various models have shown that exclusive inactivation of MBD2 does not affect the body function. *MBD2*-knockout mice demonstrate disrupted maternal behavior while nurturing and feeding their offspring [120, 124]. Aside from this, MBD2 removal does not cause any pronounced neurological deviations. Therefore, we can expect MBD2 inhibition to not cause severe side-effects in humans. Thus, the methyl-DNA binding repressor activity of MBD2 may be used for hemoglobin level restoration in patients with sickle-cell disease and beta thalassemia. However, inactivating the methyl-DNA binding protein case cited above is not always necessary. For instance, mutations in or inactivation of

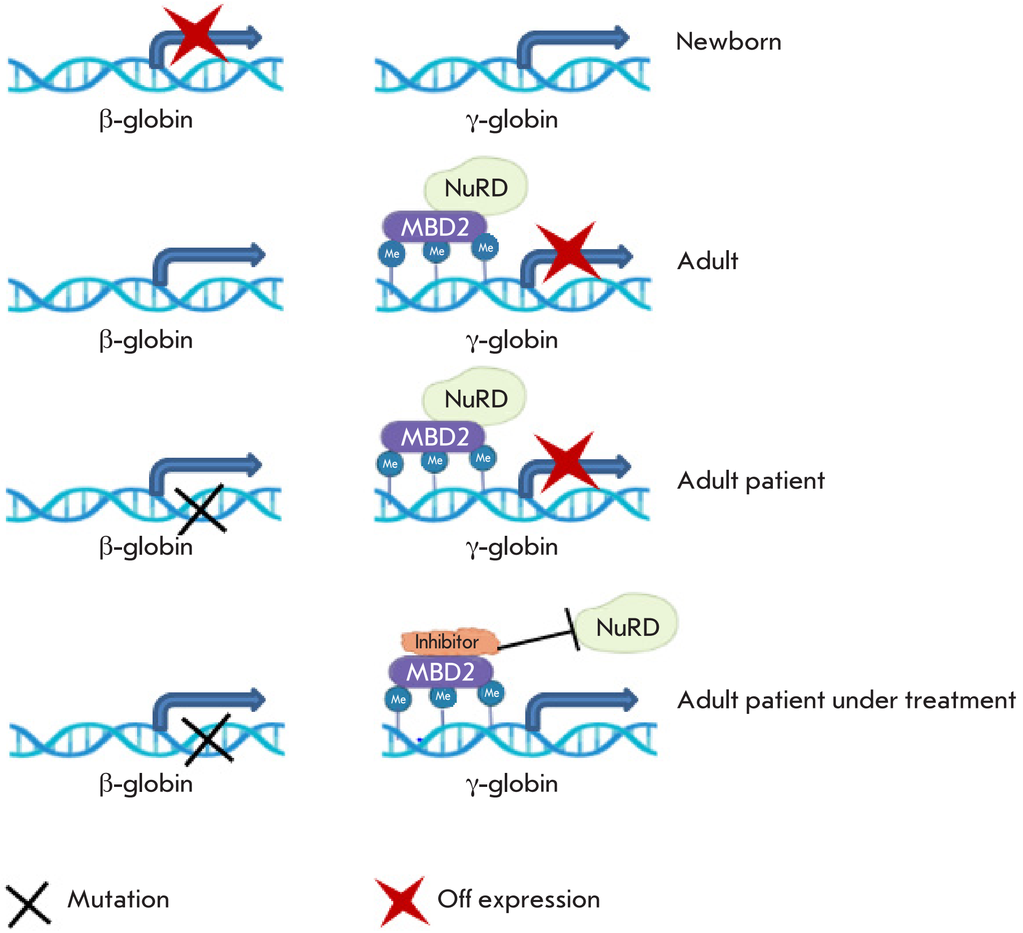


Fig. 5. Model of the functional significance of the interaction between the MBD2 protein and the NuRD repression complex in the regulation of the gamma globin gene in beta-telassemia [122, 123]

the methyl-DNA binding protein MeCP2 lead to Rett syndrome development. MeCP2 knockout in mice, similarly to mutations in this gene in humans, causes neurological changes. Notably, changes occurring in nerve cells due to MeCP2 removal or mutation are reversible [125]. The MeCP2 mutations identified in patients with Rett syndrome include, among others, point mutations causing MeCP2 degradation but not affecting the structure of its DNA-binding and repressor domains [126]. When stabilized, this protein can still fulfill its functions [127]. A search for small molecules binding to MeCP2 ubiquitination sites could make it possible to prevent its ubiquitination, with subsequent degradation, and restore the protein's functional activity.

Thus, the search for and characterization of the binding sites in methyl-DNA binding proteins are necessary for the identification of potential targets whose activity is regulated by DNA methylation and the formation of repressor complexes. Further analysis of the various pathogenetic conditions associated with the target genes of methyl-DNA binding proteins allows us to consider methyl-DNA binding proteins as targets for therapy, while investigation of the mutations in methyl-DNA binding proteins makes it possible to understand when functional changes caused by mutations can be compensated, and when that is impossible.

CONCLUSIONS

DNA methylation is a regulatory element critical to gene expression, genome stabilization, inactivation of repeating sequences, establishment of imprinting, and X-inactivation. Advanced genome-wide sequencing methods allowed us to determine the DNA methylation pattern across the whole genome, including various repeating sequences. It opened new opportunities in terms of the identification and characterization of regulatory elements whose activity may be disrupted by various pathogenetic conditions. As of now, a lot of TFs participating in methylation onset and maintenance, demethylation, or interpretation of methylated DNA have been discovered. Methylation can facilitate TF attraction or interfere with it; i.e., the DNA methylation level affects selection of the protein factors interacting with DNA and alternating between attraction of transcription activators and repressors. Discovery of new DNA methylation-dependent factors and investigation of the activating and repressor complexes they are included in allow us to consider these factors as new therapy targets to be manipulated to achieve a more nuanced effect compared to genome-wide inhibition of DNA methylation. Thus, the study of new methyl-DNA sensitive proteins could make it possible to identify new approaches and therapeutic targets for the management of various pathogenetic conditions associated with DNA methylation onset and regulatory changes. ●

This work was supported by the Russian Foundation for Basic Research (project No. 19-29-04139) (Chapters 3–4) and the Russian Scientific Foundation (project No.19-74-30026) (Chapters 1–2).

Authors have no conflict of interest to declare.

REFERENCES

1. Lister R., Mukamel E.A., Nery J.R., Urich M., Puddifoot C.A., Johnson N.D., Lucero J., Huang Y., Dwork A.J., Schultz M.D., et al.// Science. 2013. V. 341. № 6146. 1237905.
2. Gershman A., Sauria M.E.G., Guitart X., Vollger M.R., Hook P.W., Hoyt S.J., Jaun M., Shumate A., Razaghi R., Koren S., et al. // Science. 2022. V. 376. № 6588. eabj5089.
3. Yagi M., Kabata M., Tanaka A., Ukai T., Ohta S., Nakabayashi K., Shimizu M., Hata K., Meissner A., Yamamoto T., et al. // Nat. Commun. 2020. V. 11. № 1. P. 3199.
4. Kato Y., Kaneda M., Hata K., Kumaki K., Hisano M., Kohara Y., Okano M., Li E., Nozaki M., Sasaki H. // Hum. Mol. Genet. 2007. V. 16. № 19. P. 2272–2280.
5. Dodd T., Yan C., Kossmann B.R., Martin K., Ivanov I. // Proc. Natl. Acad. Sci. USA. 2018. V. 115. № 23. P. 5974–5979.
6. Shi D.-Q., Ali I., Tang J., Yang W.-C. // Front. Genet. 2017. V. 8. P. 100.
7. Velasco G., Grillo G., Touleimat N., Ferry L., Ivkovic I., Ribierre F., Deleuze J.-F., Chantalat S., Picard C., Francastel C. // Hum. Mol. Genet. 2018. V. 27. № 14. P. 2409–2424.
8. Sun Z., Wu Y., Ordog T., Baheti S., Nie J., Duan X., Hojo K., Kocher J.-P., Dyck P.J., Klein C.J. // Epigenetics. 2014. V. 9. № 8. P. 1184–1193.
9. Wang X., Zhao C., Zhang C., Mei X., Song J., Sun Y., Wu Z., Shi W. // Cell Commun. Signal. 2019. V. 17. № 1. P. 1183–1193.
10. Salameh Y., Bejaoui Y., Hajj E.N. // Front. Genet. 2020. V. 11. P. 171.
11. Nishiyama A., Nakanishi M. // Trends Genet. 2021. V. 37. № 11. P. 1012–1027.
12. Hwang J.-Y., Zukin R.S. // Curr. Opin. Neurobiol. 2018. V. 48. P. 193–200.
13. Day J.J., Sweatt D.J. // Nat. Neurosci. 2010. P. 1319–1323.
14. Rollins R.A., Haghighi F., Edwards J.R., Das R., Zhang M.Q., Ju J., Bestor T.H. // Genome Res. 2006. V. 16. № 2. P. 157–163.
15. Toubiana S., Larom G., Smoom R., Duszynski R.J., Godley L.A., Francastel C., Velasco G., Selig S. // Hum. Mol. Genet. 2020. V. 29. № 19. P. 3197–3210.
16. Toubiana S., Velasco G., Chityat A., Kaindl A.M., Hershtig N., Tzur-Gilat A., Francastel C., Selig S. // Hum. Mol. Genet. 2018. V. 27. № 20. P. 3568–3581.
17. Rajshekar S., Yao J., Arnold P.K., Payne S.G., Zhang Y., Bowman T.V., Schmitz R.J., Edwards J.R., Goll M. // eLife. 2018. V. 7. e39658. doi: 10.7554/elife.39658
18. Scelfo A., Fachinetti D. // Cells. 2019. V. 8. № 8. P. 912.
19. Guenther M.G., Levine S.S., Boyer L.A., Jaenisch R., Young R.A. // Cell. 2007. V. 130. № 1. P. 77–88.
20. Mohn F., Weber M., Rebhan M., Roloff T.C., Richter J., Stadler M.B., Bibel M., Schübeler D. // Mol. Cell. 2008. V. 30. № 6. P. 755–766.
21. Sproul D., Nestor C., Culley J., Dickson J.H., Dixon M., Harrison D.J., Meehan R.R., Sims A.H., Ramsahoye B.H. // Proc. Natl. Acad. Sci. USA. 2011. V. 108. P. 4364–4369.
22. Wan J., Oliver V.F., Wang G., Zhu H., Zack D.J., Merbs S.L., Qian J. // BMC Genomics. 2015. V. 16. № 1. P. 49.
23. Illingworth R.S., Gruenewald-Schneider U., Webb S., Kerr A.R.W., James K.D., Turner D.J., Smith C., Harrison D.J., Andrews R., Bird A.P., et al. // PLoS Genet. 2010. V. 6. № 9. e1001134.
24. Jeziorska D.M., Murray R.J.S., De Gobbi M., Gaentzsch R., Garrick D., Ayyub H., Chen T., Li E., Telenius J., Lynch M., et al. // Proc. Natl. Acad. Sci. USA. 2017. V. 114. № 36. P. E7526–E7535.
25. Wu H., Coskun V., Tao J., Xie W., Ge W., Yoshikawa K., Li E., Zhang Y., Sun Y.E. // Science. 2010. V. 329. № 5990. P. 444–448.
26. Bell J.S.K., Vertino P.M. // Epigenetics. 2017. V. 12. № 6.

P. 449–464.

27. Pachano T., Sánchez-Gaya V., Ealo T., Mariner-Faulí M., Bleckwehl T., Asenjo H.G., Respuela P., Cruz-Molina S., Muñoz-San Martín M., Haro E., et al. // Nat. Genet. 2021. V. 53. № 7. P. 1036–1049.
28. Zhang Y., Jurkowska R., Soeroes S., Rajavelu A., Dhayalan A., Bock I., Rathert P., Brandt O., Reinhardt R., Fischle W., et al. // Nucl. Acids Res. 2010. V. 38. № 13. P. 4246–4253.
29. Otani J., Nankumo T., Arita K., Inamoto S., Ariyoshi M., Shirakawa M. // EMBO Rep. 2009. V. 10. № 11. P. 1235–1241.
30. Zhang Z.-M., Liu S., Lin K., Luo Y., Perry J.J., Wang Y., Song J. // J. Mol. Biol. 2015. V. 427. № 15. P. 2520–2531.
31. Long H.K., Blackledge N.P., Klose R.J. // Biochem. Soc. Trans. 2013. V. 41. № 3. P. 727–740.
32. Brown D.A., Di Cerbo V., Feldmann A., Ahn J., Ito S., Blackledge N.P., Nakayama M., McClellan M., Dimitrova E., Turberfield A.H., et al. // Cell Rep. 2017. V. 20. № 10. P. 2313–2327.
33. Clouaire T., Webb S., Skene P., Illingworth R., Kerr A., Andrews R., Lee J.-H., Skalnik D., Bird A. // Genes Dev. 2012. V. 26. № 15. P. 1714–1728.
34. Costa Y., Ding J., Theunissen T.W., Faiola F., Hore T.A, Shliaha P.V., Fidalgo M., Saunders A., Lawrence M., Dietmann S., et al. // Nature. 2013. V. 495. P. 370–374.
35. Deplus R., Delatte B., Schwinn M.K., Defrance M., Méndez J., Murphy N., Dawson M.A., Volkmar M., Putmans P., Calonne E., et al. // EMBO J. 2013. V. 32. № 5. P. 645–655.
36. Sardina J.L., Collombet S., Tian T.V., Gómez A., Di Stefano B., Berenguer C., Brumbaugh J., Stadhouders R., Segura-Morales C., Gut M., et al. // Cell Stem Cell. 2018. V. 23. № 6. P. 727–741.e9.
37. Fernandez Garcia M., Moore C.D., Schulz K.N., Alberto O., Donague G., Harrison M.M., Zhu H., Zaret K.S. // Mol. Cell. 2019. V. 75. № 5. P. 921–932.e6.
38. Michael A.K., Grand R.S., Isbel L., Cavadini S., Kozicka Z., Kempf G., Bunker R. D., Schenk A.D., Graff-Meyer A., Pathare G.R., et al. // Science. 2020. V. 368. № 6498. P. 1460–1465.
39. Donovan B.T., Chen H., Jipa C., Bai L., Poirier M.G. // Elife. 2019. V. 8. № 7. e43008.
40. VandenBosch L.S., Wohl S.G., Wilken M.S., Hooper M., Finkbeiner C., Cox K., Chipman L., Reh T.A. // Sci. Rep. 2020. V. 1. P. 13615.
41. Metzakopian E., Bouhali K., Alvarez-Saavedra M., Whitsett J.A., Picketts D.J., Ang S.-L. // Development. 2015. V. 142. № 7. P. 1315–1324.
42. Yin Y., Morgunova E., Jolma A., Kaasinen E., Sahu B., Khund-Sayeed S., Das P.K., Kivioja T., Dave K., Zhong F., et al. // Science. 2017. V. 356. № 6337. aaj2239.
43. Lu F., Liu Y., Jiang L., Yamaguchi S., Zhang Y. // Genes Dev. 2014. V. 28. № 19. P. 2103–2119.
44. Arab K., Karaulanov E., Musheev M., Trnka P., Schäfer A., Grummt I., Niehrs C. // Nat. Genet. 2019. V. 51. P. 217–223.
45. Marsico G., Chambers V.S., Sahakyan A.B., McCauley P., Boutell J.M., Antonio M.D., Balasubramanian S. // Nucl. Acids Res. 2019. V. 47. № 8. P. 3862–3874.
46. Chashchina G.V., Beniaminov A.D., Kaluzhny D.N. // Biochemistry (Moscow). 2019. V. 84. № 5. P. 562–569.
47. Chambers V.S., Marsico G., Boutell J.M., Di Antonio M., Smith G.P., Balasubramanian S. // Nat. Biotechnol. 2015. V. 33. № 8. P. 877–881.
48. Isaakova E., Varizhuk A., Pozmogova G. // Signif. Bioeng. Biosci. 2018. V. 1. № 3. P. 1–7.
49. Chashchina G.V., Shchyolkina A.K., Kolosov S.V., Beniaminov A.D., Kaluzhny D.N. // Front. Microbiol. 2021. V. 12. P. 647851.
50. Wu F., Niu K., Cui Y., Li C., Lyu M., Ren Y., Chen Y., Deng H., Huang L., Zheng S., et al. // Commun. Biol. 2021. V. 4. № 1. P. 98.
51. Stevens A.J., de Jong L., Kennedy M.A. // Int. J. Mol. Sci. 2022. V. 23. № 5. P. 2407.
52. Hasegawa H., Sasaki I., Tsukakoshi K., Ma Y., Nagasawa K., Numata S., Inoue Y., Kim Y., Ikebukuro K. // Int. J. Mol. Sci. 2021. V. 22. № 23. P. 13159.
53. Mao S.-Q., Ghanbarian A.T., Spiegel J., Martínez Cuesta S., Beraldi D., Di Antonio M., Marsico G., Hänsel-Hertsch R., Tannahill D., Balasubramanian S. // Nat. Struct. Mol. Biol. 2018. V. 25. P. 951–957.
54. Cree S.L., Fredericks R., Miller A., Pearce F.G., Filichev V., Fee C., Kennedy M. A. // FEBS Lett. 2016. V. 590. № 17. P. 2870–2883.
55. Jara-Espejo M., Line S.R. // FEBS J. 2020. V. 287. № 3. P. 483–495.
56. Haggerty C., Kretzmer H., Riemenschneider C., Kumar A.S., Mattei A.L., Bailly N., Gottfreund J., Giesselmann P., Weigert R., Brändl B., et al. // Nat. Struct. Mol. Biol. 2021. V. 28. № 7. P. 594–603.
57. Tian H.-P., Lun S.-M., Huang H.-J., He R., Kong P.-Z., Wang Q.-S., Li X.-Q., Feng Y.-M. // J. Biol. Chem. 2015. V. 290. № 31. P. 19173–19183.
58. Lienert F., Wirbelauer C., Som I., Dean A., Mohn F., Schübeler D. // Nat. Genet. 2011. V. 43. № 11. P. 1091–1097.
59. Braccioli L., de Wit E. // Essays Biochem. 2019. V. 63. № 1. P. 157–165.
60. Schoenherr C.J., Levorse J.M., Tilghman S.M. // Nat. Genet. 2003. V. 33. № 1. P. 66–69.
61. Fedoriw A.M., Stein P., Svoboda P., Schultz R.M., Bartolomei M.S. // Science. 2004. V. 303. № 5655. P. 238–240.
62. Freschi A., Del Prete R., Pignata L., Cecere F., Manfrevola F., Mattia M., Cobellis G.,. Sparago A., Bartolomei M.S., Riccio A., et al. // Hum. Mol. Genet. 2021. V. 30. № 16. P. 1509–1520.
63. Damaschke N.A., Gawdzik J., Avilla M., Yang B., Svaren J., Roopra A., Luo J.-H., Yu Y.P., Keles S., Jarrard D.F., et al. // Clin. Epigenetics. 2020. V. 12. № 1. P. 80.
64. Luo X., Zhang T., Zhai Y., Wang F., Zhang S., Wang G. // Front. Genet. 2021. V. 12. P. 639461.
65. Hashimoto H., Wang D., Horton J.R., Zhang X., Corces V.G., Cheng X. // Mol. Cell. 2017. V. 66. № 5. P. 711–720.
66. Stadler M.B., Murr R., Burger L., Ivanek R., Lienert F., Schöler A., van Nimwegen E., Wirbelauer C., Oakeley E.J., Gaidatzis D., et al. // Nature. 2011. V. 480. P. 490–495.
67. Reshef Y.A., Finucane H.K., Kelley D.R., Gusev A., Kotliar D., Ulirsch J.C., Hormozdiari F., Nasser J., O'Connor L., van de Geijn B., et al. // Nat. Genet. 2018. V. 50. № 10. P. 1483–1493.
68. Kurukuti S., Tiwari V.K., Tavoosidana G., Pugacheva E., Murrell A., Zhao Z., Lobanenkov V., Reik W., Ohlsson R. // Proc. Natl. Acad. Sci. USA. 2006. V. 103. № 28. P. 10684–10689.
69. Schmitz K.-M., Mayer C., Postepska A., Grummt I. // Genes Dev. 2010. V. 24. № 20. P. 2264–2269.
70. Baubec T., Colombo D.F., Wirbelauer C., Schmidt J., Burger L., Krebs A.R., Akalin A., Schubeler D. // Nature. 2015. V. 520. P. 243–247.
71. Weinberg D.N., Papillon-Cavanagh S., Chen H., Yue Y., Chen X., Rajagopalan K.N., Horth C., McGuire J.T., Xu X.,

Nikbakht H. // Nature. 2019. V. 573. P. 281–286.
72. Li Y., Chen X., Lu C. // EMBO Rep. 2021. V. 22. № 5. e51803.
73. Hervouet E., Vallette F.M., Cartron P.-F. // Epigenetics. 2009. V. 4. № 7. P. 487–499.
74. Sato N., Kondo M., Arai K.-I. // Biochem. Biophys. Res. Commun. 2006. V. 344. № 3. P. 845–851.
75. Gu P., Le Menuet D., Chung A.C.-K., Cooney A.J. // Mol. Cell. Biol. 2006. V. 26. № 24. P. 9471–9483.
76. Tullius T., Parker S. // Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature. 2013. V. 152. № 1–2.
77. Hudson N.O., Buck-Koehntop B.A. // Molecules. 2018. V. 23. № 10. P. 2555.
78. Prokhortchouk A., Hendrich B., Jørgensen H., Ruzov A., Wilm M., Georgiev G., Bird A., Prokhortchouk E. // Genes Dev. 2001. V. 15. № 13. P. 1613–1618.
79. Filion G.J.P., Zhenilo S., Salozhin S., Yamada D., Prokhortchouk E., Defossez P.-A. // Mol. Cell Biol. 2006. V. 26. № 1. P. 169–181.
80. Daniel J.M., Spring C.M., Crawford H.C., Reynolds A.B., Baig A. // Nucl. Acids Res. 2002. V. 30. № 13. P. 2911–2919.
81. Nikolova E.N., Stanfield R.L., Dyson H.J., Wright P.E. // Biochemistry. 2018. V. 57. № 14. P. 2109–2120.
82. Liu Y., Toh H., Sasaki H., Zhang X., Cheng X. // Genes Dev. 2012. V. 26. № 21. P. 2374–2379.
83. Zhenilo S., Deyev I., Litvinova E., Zhigalova N., Kaplun D., Sokolov A., Mazur A., Prokhortchouk E. // Cell Death Differ. 2018. V. 25. № 11. P. 1938–1951.
84. Yoon H.-G., Chan D.W., Reynolds A.B., Qin J., Wong J. // Mol. Cell. 2003. V. 12. № 3. P. 723–734.
85. Raghav S.K., Waszak S.M., Krier I., Gubelmann C., Isakova A., Mikkelsen T.S., Deplancke B. // Mol. Cell. 2012. V. 46. № 3. P. 335–350.
86. Bohne F., Langer D., Martiné U., Eider C.S., Cencic R., Begemann M., Elbracht M., Bülow L., Eggermann T., Zechneret U., et al. // Clin. Epigenet. 2016. V. 8. P. 47.
87. Prokhortchouk A., Sansom O., Selfridge J., Caballero I.M., Salozhin S., Aithozhina D., Cerchietti L., Guo Meng F., Augenlicht L.H., Mariadason J.M., et al. // Mol. Cell. Biol. 2006. V. 26. № 1. P. 199–208.
88.Kaplun D.S., Fok R.E., Korostina V.S., Prokhortchouk E.B., Zhenilo S.V. // Biochemistry (Moscow). 2019. V. 84. № 3. P. 283–290.
89. Kaplun D., Starshin A., Sharko F., Gainova K., Filonova G., Zhigalova N., Mazur A., Prokhortchouk E., Zhenilo S., et al. // Int. J. Mol. Sci. 2021. V. 22. № 14. P. 7587.
90. Shi X., Du Y., Li S., Wu H. // Int. J. Mol. Sci. 2022. V. 23. № 7. P. 3639.
91. Kagey M.H., Melhuish T.A., Wotton D. // Cell. 2003. V. 113. № 1. P. 127–137.
92. Zhenilo S., Kaplun D., Prokhortchouk E. // FEBS Open Bio. 2018. V. 8 (S1). P. 134.
93. Lee B., Muller M.T. // Biochem. J. 2009. V. 421. № 3. P. 449–461.
94. Steinacher R., Barekati Z., Botev P., Kuśnierczyk A., Slupphaug G., Schär P. // EMBO J. 2019. V. 38. № 1. e99242.
95. Quenneville S., Verde G., Corsinotti A., Kapopoulou A., Jakobsson J., Offner S., BaglivoI., Pedone P.V., Grimaldi G., Riccio A., et al. // Mol. Cell. 2011. V. 44. № 3. P. 361–372.
96. Riso V., Cammisa M., Kukreja H., Anvar Z., Verde G., Sparago A., Acurzio B., Lad S., Lonardo E., Sankar A., et al. // Nucl. Acids Res. 2016. V. 44. № 17. P. 8165–8178.
97. Shi H., Strogantsev R., Takahashi N., Kazachenka A., Lorincz M.C., Hemberger M., Ferguson-Smith A.C. // Epigenetics Chromatin. 2019. V. 12. № 1. P. 49.
98. Mancini M., Magnani E., Macchi F., Bonapace I.M. // Nucl. Acids Res. 2021. V. 49. № 11. P. 6053–6068.
99. Qin W., Wolf P., Liu N., Link S., Smets M., La Mastra F., Forné I., Pichler G., Hörl D., Fellinger K., et al. // Cell Res. 2015. V. 25. № 8. P. 911–929.
100. Beck A., Trippel F., Wagner A., Joppien S., Felle M., Vokuhl C., Schwarzmayr T., Strom T.M., von Schweinitz D., Längst G., et al. // Clin. Epigenetics. 2018. V. 10. P. 27.
101. Tian Y., Paramasivam M., Ghosal G., Chen D., Shen X., Huang Y., Akhter S., Legerski R., Chen J., Seidman M.M., et al. // Cell Rep. 2015. V. 10. № 12. P. 1957–1966.
102. Baubec T., Ivánek R., Lienert F., Schübeler D. // Cell. 2013. V. 153. P. 480–492.
103. Jin Y., Su K., Kong H.E., Ma W., Wang Z., Li Y., R. Li, Allen E. G., Wu H., Jin P. // Hum. Mol. Genet. 2022. ddac189.
104. Allan A.M., Liang X., Luo Y., Pak C., Li X., Szulwach K.E., Chen D., Jin P., Zhao X. // Hum Mol. Genet. 2008. V. 17. № 13. P. 2047–2057.
105. Li X., Barkho B.Z., Luo Y., Smrt R.D., Santistevan N.J., Liu C., Kuwabara T., Gage F. H., Zhao X. // J. Biol. Chem. 2008. V. 283. № 41. P. 27644–27652.
106. Katayama S., Andou M. // Biochem. Biophys. Res. Commun. 2021. V. 581. P. 20–24.
107. O'Geen H., Bates S.L., Carter S.S., Nisson K.A., Halmai J., Fink K.D., Rhie S.K., Farnham P.J., Segal D.J. // Epigenetics Chromatin. 2019. V. 12. P. 26.
108. Choudhury S.R., Cui Y., Lubecka K., Stefanska B., Irudayaraj J. // Oncotarget. 2016. V. 7. № 29. P. 46545–46556.
109. Kang J.G., Park J.S., Ko J.-H., Kim Y.-S. // Sci. Rep. 2019. V. 9. № 1.. P. 11960.
110. Devesa-Guerra I., Morales-Ruiz T., Pérez-Roldán J., Parrilla-Doblas J.T., Dorado-León M., García-Ortiz M.V., et al. // J. Mol. Biol. 2020. V. 432. № 7. P. 2204–2216.
111. Ciechomska M., Roszkowski L., Maslinski W. // Cells. 2019. V. 8. № 9. P. 953.
112. Jung G., Hernández-Illán E., Moreira L., Balaguer F., Goel A. // Nat. Rev. Gastroenterol. Hepatol. 2020. V. 17. № 2. P. 111–130.
113. Müller D., Győrffy B. // Biochim. Biophys. Acta Rev. Cancer. 2022. V. 1877. № 3. 188722.
114. Taryma-Leśniak O., Sokolowska K.E., Wojdacz T.K. // Clin Epigenetics. 2020. V. 12. № 1. P. 107.
115. Brocks D., Schmidt C.R., Daskalakis M., Jang H.S., Shah N.M., Li D., Li J., Zhang B., Hou Y., Laudato S., et al. // Nat. Genet. 2017. V. 49. № 7. P. 1052–1060.
116. Graça I., Sousa E.J., Baptista T., Almeida M., Ramalho-Carvalho J., Palmeira C., Henrique R., Jerónimo C. // Curr. Pharm. Des. 2014. V. 20. № 11. P. 1803–1811.
117. Segura-Pacheco B., Perez-Cardenas E., Taja-Chayeb L., Chavez-Blanco A., Revilla-Vazquez A., Benitez-Bribiesca L., Duenas-González A.. // J. Transl. Med. 2006. V. 4. P. 32.
118. Ansari I., Chaturvedi A., Chitkara D., Singh S. // Semin. Cancer Biol. 2022. V. 83. P. 570–583.
119. Hanzawa N., Hashimoto K., Yuan X., Kawahori K., Tsujimoto K., Hamaguchi M., Tanaka T., Nagaoka Y., Nishina H., Morita S., et al. // Sci. Rep. 2020. V. 10. № 1. P. 5181.
120. Wood K.H., Johnson B.S., Welsh S.A., Lee J.Y., Cui Y., Krizman E., Brodkin E.S., Blendy J.A., Robinson M.B., Bartolomei M.S., et al. // Epigenomics. 2016. V. 8. № 4. P. 455–473.

121. Kulikov A.V., Korostina V.S., Kulikova E.A., Fursenko D.V., Akulov A.E., Moshkin M.P., Prokhortchouk E.B. // Behav. Brain Res. 2016. V. 297. P. 76–83.
122. Gnanapragasam M.N., Scarsdale J.N., Amaya M.L., Webb H.D., Desai M.A., Walavalkar N.M., Wang S.Z., Zhu S.Z., Ginder G.D., Williams Jr., D.C. // Proc. Natl. Acad. Sci. USA. 2011. V. 108. № 18. P. 7487–7492.
123. Yu X., Azzo A., Bilinovich S.M., Li X., Dozmorov M., Kurita R., Nakamura Y., Williams Jr., D.C., Ginder G.D. // Haematologica. 2019. V. 104. № 12. P. 2361–2371.
124. Hendrich B., Guy J., Ramsahoye B., Wilson V.A., Bird A. // Genes Dev. 2001. V. 15. № 6. P. 710–723.
125. Robinson L., Guy J., McKay L., Brockett E., Spike R.C., Selfridge J., De Sousa D., Merusi C., Riedel G., Bird A., et al. // Brain. 2012. V. 135. P. 9. P. 2699–2710.
126. Scarsdale J.N., Webb H.D., Ginder G.D., Williams D.C., Jr. // Nucl. Acids Res. 2011. V. 39. № 15. P. 6741–6752.
127. Ghosh R.P., Horowitz-Scherer R.A., Nikitina T., Gierasch L.M., Woodcock C.L. // J. Biol. Chem. 2008. V. 283. № 29. P. 20523–20534.

Flow-Seq Method: Features and Application in Bacterial Translation Studies

E. S. Komarova[1], O. A. Dontsova[2,3,4,5], D. V. Pyshnyi[6], M. R. Kabilov[6*], P. V. Sergiev[1,2,3,4*]
[1]Institute of Functional Genomics, Lomonosov Moscow State University, Moscow, 119234 Russia
[2]Department of Chemistry, Lomonosov Moscow State University, Moscow, 119234 Russia
[3]Skolkovo Institute of Science and Technology, Moscow, 121205 Russia
[4]Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, 119234 Russia
[5]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow 117437 Russia
[6]Institute of Chemical Biology and Fundamental Medicine, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, 630090 Russia
*E-mail: petya@genebee.msu.ru, kabilov@niboch.nsc.ru

Received October 04, 2022; in final form, November 11, 2022
DOI: 10.32607/actanaturae.11820

ABSTRACT The Flow-seq method is based on using reporter construct libraries, where a certain element regulating the gene expression of fluorescent reporter proteins is represented in many thousands of variants. Reporter construct libraries are introduced into cells, sorted according to their fluorescence level, and then subjected to next-generation sequencing. Therefore, it turns out to be possible to identify patterns that determine the expression efficiency, based on tens and hundreds of thousands of reporter constructs in one experiment. This method has become common in evaluating the efficiency of protein synthesis simultaneously by multiple mRNA variants. However, its potential is not confined to this area. In the presented review, a comparative analysis of the Flow-seq method and other alternative approaches used for translation efficiency evaluation of mRNA was carried out; the features of its application and the results obtained by Flow-seq were also considered.

KEYWORDS Flow-seq, NGS, high-throughput sequencing, flow cytometry, translation, bacteria.

ABBREVIATIONS TIR – translation initiation region; RBS – ribosome binding site; SD – Shine–Dalgarno sequence; 5' UTR – 5' untranslated region; ORF – open reading frame; NGS – next-generation sequencing; Flow-seq – flow cytometry and next-generation sequencing.

INTRODUCTION

Translation is the key process in the vital activity of all organisms, during which proteins are synthesized in cells using a macromolecular ribonucleoprotein complex known as the ribosome. It decodes the information in mRNA and translates it into the sequence of amino acids that form the protein [1]. Moreover, not only does mRNA participate in this process as a passive information carrier, but it also predetermines the translation efficiency [2].

The 5' untranslated region (5' UTR) of mRNA is one of the sites responsible for its translation efficiency (*Fig. 1A*) [3]. The 5' UTR contains the ribosome binding site (RBS) carrying the Shine–Dalgarno (SD) sequence [4–13] complementary to the 3' terminus of 16S rRNA in canonical mRNAs [14, 15]. To ensure high efficiency of the protein synthesis, the SD sequence needs to be located at an optimal distance from the start codon and have an optimal length [16–18]. Sometimes a single 5' UTR can carry several Shine–Dalgarno sequences [2, 17]. For efficient translation, the translation initiation region (TIR) needs to be either fully single-stranded or folded into the secondary structure that can be easily disturbed [19–22]. Other elements capable of affecting the translation efficiency are known, such as the adenine- and uracil-rich (AU-rich) mRNA region that the ribosomal protein bS1 interacts with [23–25], as well as the initial portion of the coding region immediately downstream of the start codon [26–28]. The 5' UTRs of efficiently translated mRNAs are characterized by low abundance of cytidine residues and the presence of purine repeats (AG repeats) [2].

Today, there are various methods that allow one to study the functional significance of individual mRNA sites for protein synthesis. These methods involve site-directed mutagenesis [29] or randomization [30, 31] of 5' UTR motifs (usually upstream of the fluorescent protein gene), and assessment of its fluorescence intensity *in vitro* (or *in vivo*), which is indicative of translation efficiency. The *in silico* thermodynamic simulations [18, 32–36], which estimate the strength of molecular interactions between the 30S complex and the mRNA transcript and predict the translation initiation rate, can be used to determine the values related to the translation efficiency. The simulation results can be selectively verified experimentally using reporter constructs. The emergence of the flow cytometry method has made it possible to simultaneously assess different parameters of a large number of cells *in vivo* and isolate individual fractions based on the similarity of certain parameters (e.g., according to the expression level of the fluorescent protein gene) [37]. Advancements in next-generation sequencing (NGS) have contributed to the development of novel, comprehensive approaches to genome research and to the determination of the genotype–phenotype correlation (e.g., whole genome sequencing, sequencing of plasmid DNA libraries, RNA sequencing for single-cell transcriptome profiling and isolation of efficiently translated mRNA, as well as ChIP sequencing for identifying the binding sites of DNA-associated proteins) [38, 39].

THE VARIETY OF APPROACHES TO STUDYING THE ROLE OF 5' UTRS IN TRANSLATION EFFICIENCY

Comprehensive analysis of *E. coli* genes has shown that most mRNAs carry the Shine–Dalgarno (SD) sequence (*Fig. 1B*), which was discovered in several bacterial mRNAs in the 1970s [4] and is essential for efficient translation initiation [16–18]. The SD sequence is the best studied regulatory element. It resides 5–8 nucleotides upstream of the start codon (or 8–11 nucleotides when starting counting from the central G base in the SD sequence [7]) and acts as a binding site to the bacterial 30S subunit, unlike in the eukaryotic ribosome, which binds to the 5' terminus of mRNA for scanning initiation [6]. Different *E. coli* mRNAs contain SD sequences of different lengths, varying between four and eight nucleotides. The most plausible composition of the SD sequence is agGa.

The dependence between the protein synthesis efficiency and the length of the SD sequence and its distance from the start codon was studied using various methods (e.g., using a dual genetically engineered construct (*Fig. 1C*) carrying the genes of two fluorescent proteins, where one of the proteins, RFP (red fluorescent protein), was an internal control and the other one, CER (cyan fluorescent protein), acted as a sensor of the effects associated with variations in the mRNA 5' UTRs) [17]. The ratio between the measured fluorescence intensities of the two proteins (CER/RFP) *in vivo* was calculated, making it possible to neutralize the effects caused by the bacterial cell size and fluctuations in the abundance of the reporter plasmid. This approach, based on molecular cloning with the use of 16 reporter constructs with four SD sequences (2, 4, 6 and 8) of different lengths residing at different distances from the start codon of the CER protein gene (7, 10, 13 and 16) and another control construct carrying no sites complementary to the anti-SD sequence, allowed the researchers to experimentally study the effect of the SD sequence length, the distance between the SD sequence and the start codon, and their combinations on the synthesis of the CER protein. Therefore, it was demonstrated that the translation efficiency of mRNA carrying the 8-nucleotide SD sequence declines with increasing distance between the start codon and the SD sequence. For the 6-nucleotide SD sequence, the optimal distance is 10 nucleotides. The same dependence was observed for the medium-length SD sequence (four nucleotides), as in the case of a long SD sequence (eight nucleotides). For the short SD sequence (two nucleotides), the effect of distance was negligible, while the role of this SD sequence in the protein synthesis efficiency was preserved: it ensured an efficiency that was one order of magnitude greater than that when using the control construct without the SD sequence. By varying these parameters, one can change the translation level by up to four orders of magnitude, which indicates that they are important for determining the level of many proteins in the cell [17].

Numerous variants of the motif in 5' UTR produced by site-directed mutagenesis based on use of the polymerase chain reaction (PCR) can be employed to perform a rapid, and fairly simple, quantitative analysis of gene expression *in vitro*. The PCR product containing the T7 promoter sequence, the tested 5' UTR variant, and the eGFP fluorescent protein gene are directly used in the coupled transcription–translation *in vitro* system from *E. coli* cells [29]. The translation efficiency in this system can be assessed according to eGFP fluorescence intensity. This method was used to produce 54 variants of 5' UTR sequences

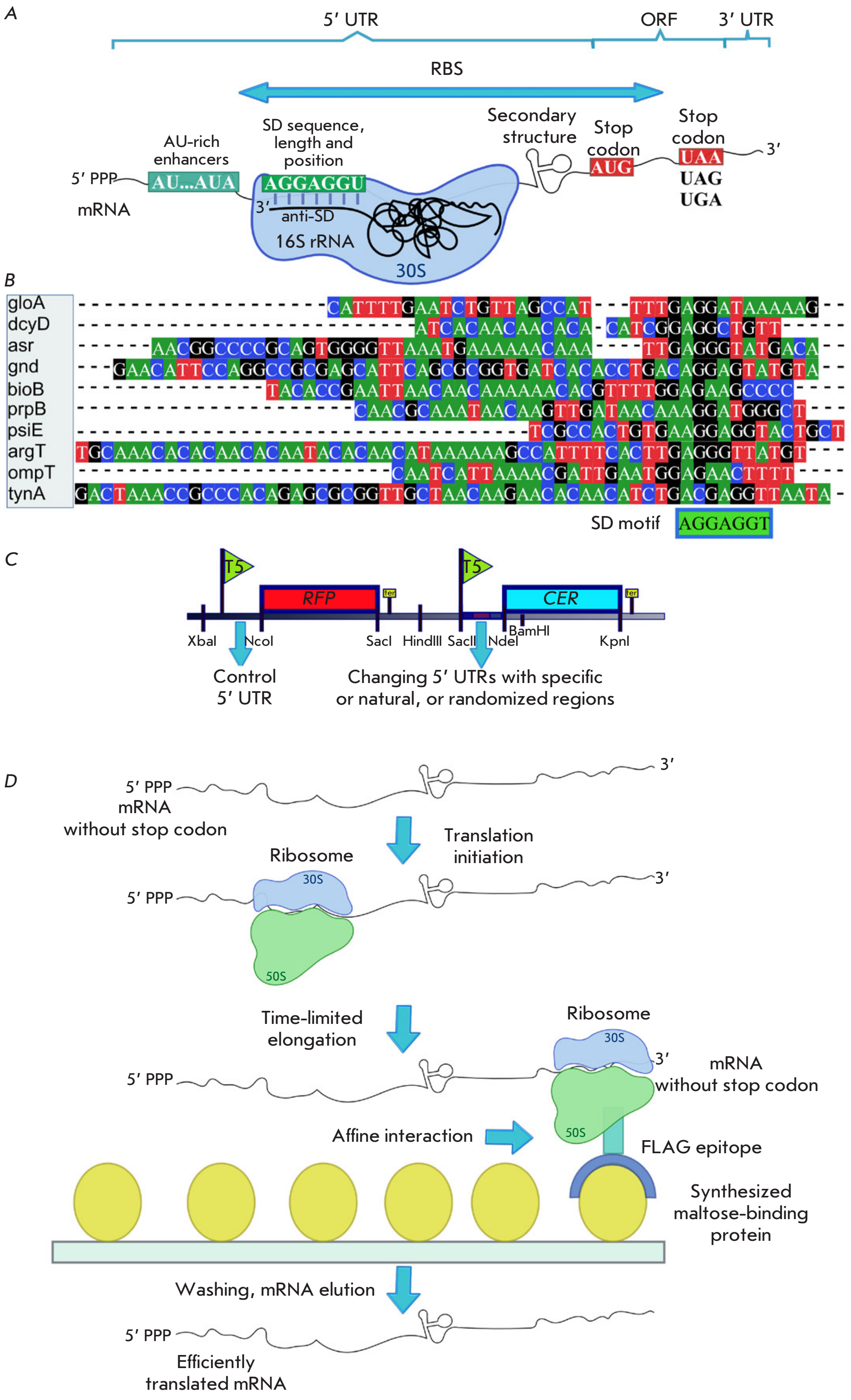

A
5′ UTR
ORF
3′ UTR
RBS
AU-rich enhancers
SD sequence, length and position
Secondary structure
Stop codon
Stop codon
5′ PPP
AU...AUA
AGGAGGU
AUG
UAA
UAG
UGA
3′
mRNA
3′
anti-SD
16S rRNA
30S
B
gloA CATTTTGAATCTGTTAGCCAT TTTGAGGATAAAAAG
dcyD ATCACAACAACACA CATCGGAGGCTGTT
asr AACGGCCCCGCAGTGGGGTTAAATGAAAAAACAAA TTGAGGGTATGACA
gnd GAACATTCCAGGCCGCGAGCATTCAGCGCGGTGATCACACCTGACAGGAGTATGTA
bioB TACACCGAATTAACAACAAAAAACACGTTTTGGAGAAGCCCC
prpB CAACGCAAATAACAAGTTGATAACAAAGGATGGGCT
psiE TCGCCACTGTGAAGGAGGTACTGCT
argT TGCAAACACACAACACAATACACAACATAAAAAAGCCATTTTCACTTGAGGGTTATGT
ompT CAATCATTAAAACGATTGAATGGAGAACTTTT
tynA GACTAAACCGCCCACAGAGCGCGGTTGCTAACAAGAACACAACATCTGACGAGGTTAATA
SD motif AGGAGGT
C
T5
RFP
ter
T5
CER
ter
XbaI
NcoI
SacI
HindIII
SacII
NdeI
BamHI
KpnI
Control 5′ UTR
Changing 5′ UTRs with specific or natural, or randomized regions
D
5′ PPP
mRNA without stop codon
3′
Translation initiation
Ribosome
30S
5′ PPP
3′
50S
Time-limited elongation
Ribosome
30S
5′ PPP
3′
mRNA without stop codon
50S
Affine interaction
FLAG epitope
Synthesized maltose-binding protein
Washing, mRNA elution
5′ PPP
Efficiently translated mRNA

Fig. 1. (*A*) – Structural features of mRNA in bacteria. 5′ and 3′ UTR – the 5′ and 3′ untranslated regions, respectively. RBS – the ribosome binding site on mRNA. ORF – the open reading frame containing the protein-coding sequence. SD and anti-SD – Shine–Dalgarno and anti-Shine–Dalgarno sequences, respectively. (*B*) – An example of 5′ UTR mRNA sequence alignment used in a large-scale analysis of untranslated gene regions with the SD motif highlighted. (*C*) – An example of a dual-reporter construct with control 5′ UTR upstream of the RFP fluorescent protein gene and a variable 5′ UTR upstream of the second CER fluorescent sensor protein gene to assess the effect of the features of the variable region on the translation efficiency. (*D*) – The scheme of affinity isolation of ribosomes with efficiently translated mRNA. Selection was carried out by limiting the time of *in vitro* translation. The mRNA contains 5′ UTR, the coding region that includes the region encoding the FLAG epitope interacting with the synthesized maltose-binding protein and TolA allowing the epitope to exit the ribosome tunnel and fold properly. There is no stop codon in the mRNA construct, so the ribosome remains on it. The drawing was executed in the Inkscape software

(18 and 36 of those having modified SD- and AU-rich sequences, respectively), which ensured a 0.1–2.0 range of relative expression levels and revealed the effects of different ribosome binding sites (RBSs) on the translation efficiency [29]. However, this pointwise approach is substantially confined to the small set of variants being tested, making it impossible to apply it to the entire variety of natural 5' UTRs lying upstream of the genes (their number in *E. coli* being ~ 4 × 10^3) [8].

An experimental system (*Fig. 1D*) [30] based on *in vitro* translation was subsequently developed, which allowed one to select the most efficiently translated mRNAs from a large sample of synthetic sequences. A model mRNA containing an 81-nucleotide 5' UTR was used for this purpose; 18 of these nucleotides, residing upstream of the start codon, were completely randomized: so, a library consisting of ~6.9 × 10^{10} different sequences was successfully obtained. The model mRNA encoded a fusion protein containing a maltose-binding domain approximately in its center and the FLAG epitope, which allowed one to perform affinity purification of the ribosomes that synthesized this fusion protein. The TolA protein fragment resided downstream of the domain used for affinity purification; this fragment acted exclusively as a spacer sufficient for affinity domain exposure from the peptide tunnel once the full-length fusion protein was synthesized. This mRNA did not carry the stop codon; therefore, it remained bound to the ribosome after the synthesis had been completed in that experiment. Therefore, mRNA could have been extracted from affinely bound ribosomes and subsequently amplified. The limited translation time was the key parameter of mRNA selection: only rapidly translated mRNAs could be affinely purified and used in the next selection round [30]. Surprisingly, 76% of the selected sequences ensuring the most rapid translation in the *in vitro* system carried no SD sequences and had C-rich short sites complementary to 16S rRNA. However, a high expression level of mRNAs with these C-rich sequences was not observed *in vivo*, which, potentially, was caused by different average ratios in the *in vitro* and *in vivo* ribosomal systems and mRNAs, which competed with C-rich RBS for ribosome binding [30]. The same experiment was conducted using a library of shorter mRNAs with a 40-nucleotide 5' UTR [31], which are the most abundant in *E. coli* mRNA [40, 41]. Next-generation sequencing and statistical tools made it possible to identify the mRNA–ribosome binding motifs. The mRNAs selected from a library with shorter 5' UTRs according to the translation rate were more likely to contain SD sequences, along with G/U-rich ones [31]. The results of this study are also indicative of the fact that the 5' UTR length affects the efficiency of protein synthesis initiation.

The sequence of mRNA 5' UTRs can be responsible for folding variations in the region upstream of the start codon. The association between the stability of the secondary structures in the TIR and the translation efficiency was confirmed by a large-scale computational analysis [19], which revealed that prokaryotic and eukaryotic genes, especially those characterized by high expression levels, tend to destabilize the mRNA secondary structure near the start codon [20]. By varying the stability (< -12 kcal/mol) of the hairpin structure carrying the RBS by site-directed mutagenesis, followed by an *in vivo* analysis of the protein yield, it was discovered that the higher the stability of the secondary structure carrying the ribosome binding site, the lower the translation efficiency. Therefore, it has been demonstrated that it is possible to vary the expression 500-fold by making a single nucleotide substitution, which stabilizes the mRNA secondary structure. As a result, translation initiation was entirely dependent on the spontaneous unfolding of the entire mRNA initiation site [21]. However, this spontaneity had to do with the fact that all the essential elements of the initiation complex were present [22]. This analysis of 12 mRNAs characterized by different levels of secondary structure stability and carrying SD sequences of different lengths (or without SD sequences) revealed that the SD sequence *per se*, the start codon, the initiator tRNA with formylated methionine, and the GTP-bound translation initiation factor 2 (IF2), in a complex with the 30S ribosomal subunit, are required for the unfolding of the mRNA secondary structure. The contribution of each individual element to the disruption of TIR mRNA folding process was assessed using the dissociation constant of the mRNA fragment carrying a 6-nucleotide SD sequence [22]. FRET analysis of the same fragment labeled with Cy3 and Cy5 at the 5' and 3' termini, in the presence of the 30S subunit and all other elements required for translation initiation, was subsequently conducted. The assessment was performed with respect to control mRNA that carried no SD sequence but whose secondary structure was characterized by a similar level of stability. The analysis revealed the significant role played by the SD sequence in the unfolding of the mRNA secondary structure. FRET analysis was shown to be highly efficient for the folded mRNA whose termini were involved in a complementary interaction between the SD and anti-SD sequences; poor efficiency of the FRET analysis was demonstrated for the unfolded mRNA [22].

The efficiency of the binding of ribosomal subunits to a particular 5' UTR mRNA sequence is assessed using the so-called toeprinting method (*Fig. 2A*). It is based on the use of fluorescent- or isotope-labeled primers complementary to the 3' terminus of mRNA. The reverse transcription reaction is performed after the assembly of the initiation complex on mRNA, followed by an electrophoretic analysis of elongated cDNA in the reaction mixture. Reverse transcriptase reaches the 5' terminus of mRNA if it is not bound to the ribosome and forms shorter products in the case when reverse transcriptase stops once it has encountered the ribosome. The ratio between long and truncated toeprints allows one to estimate the proportion of mRNAs bound to the ribosome [42, 43].

As the experimental results are acquired and methods for analyzing them are developed, bioinformatic approaches allowing one to work with large datasets start to play an increasingly important role. Translation initiation of prokaryotic mRNAs (where the SD sequence has not been detected in 5' UTR) observed in the experiments occurs independently of any interactions with the anti-SD sequence and is mediated by the ribosomal protein bS1. A bioinformatic analysis showed that the stability of the secondary structures of such 5' UTR sequences was reduced, thus facilitating the formation of the initiation complex and compensating for the lack of SD and anti-SD interactions [44, 45].

There exist the so-called prokaryotic leaderless mRNAs, which carry neither the 5' UTR nor the SD sequence. However, a large-scale *in silico* analysis of the macroevolution revealed that the number of such genes in bacteria has declined over time. The translation initiation sites of all the genes in 953 bacterial and 72 archaeal genomes have been examined and categorized into groups, according to the distance to the root (between bacteria and archaea) on the 16S rRNA-based phylogenetic tree. The average proportion of leaderless genes in each group was calculated: first, it drops rapidly and subsequently fluctuates at a low level [46].

Intense development of next-generation sequencing methods and the accumulated skills in working with the translation system have made it possible to develop the ribosome profiling (Ribo-seq) method (*Fig. 2B*), which is based on high-throughput sequencing of mRNA fragments protected by the translating ribosome [47]. This approach proved to be effective for studying gene expression, simultaneously at both the transcriptional and translational levels, including in response to various impacts [48–50]. The Ribo-seq technique provides information about the location of ribosomes on mRNA with a single-nucleotide resolution. This accuracy allows one to detect translation of mRNA sites outside of the annotated reading frame, as well as detect translation of the overlapping reading frames and semantic stop codon decoding. The translatable reading frames were identified using Ribo-seq in RNAs that had been previously considered non-coding. It was also possible to evaluate the effect of various conditions and factors on mRNA translation in cells (e.g., different environments, modifications of proteins and antibiotics) [51–56].

The extensive use of the Ribo-seq method has unearthed a number of challenges and artifacts related to the experimental methodology and data analysis [57–59]. The promising ribosome profiling technique used to study the ribosome decoding rate is characterized by infrequent high peaks in the ribosome footprint density and by long alignment gaps of the respective mRNA sequences. In order to reduce the impact of data heterogeneity, a normalization method has been elaborated. This method is efficient in the presence of heterogeneous noise and has revealed significant differences in read distribution across mRNA and the determinants of ribosome footprint frequencies in 30 publicly available ribosome profiling datasets, thereby casting doubt on the reliability of this method as an accurate representation of local ribosome density without prior quality control [57]. This observation suggests an incomplete understanding of how the protocol parameters affect the ribosome footprint density.

The most likely reason for this observation probably consists in the sequence shifting that occurs during the construction of the ribosome footprint library and its conversion into cDNA, followed by sequencing [58]. The aforementioned steps involve a number of reactions using sequence-specific enzymes such as nucleases [60]. Meanwhile, some antibiotics used to treat ribosomes prior to profiling have the same sequence specificity [61–63], which must be taken into account in experiment setting.

It has been shown using ribosome profiling in bacteria that ribosome occupancy downstream of the Shine–Dalgarno sequences occurring randomly in the coding region is significantly increased [64]. Whereas the SD sequences upstream of the start codon play a well-characterized role in translation initiation, the findings indicate that elongation is slowed down by the formation of transient base pairs between the SD motifs within the open reading frames and the anti-SD sequence in 16S rRNA, such pauses accounting for over 70% of the strong pauses throughout the genome; they are considered to be the main determinant of translational pausing in bacteria [64].

Later, the modified high-resolution Ribo-seq method was used to demonstrate that the previously observed enrichment of the ribosome occupancy at

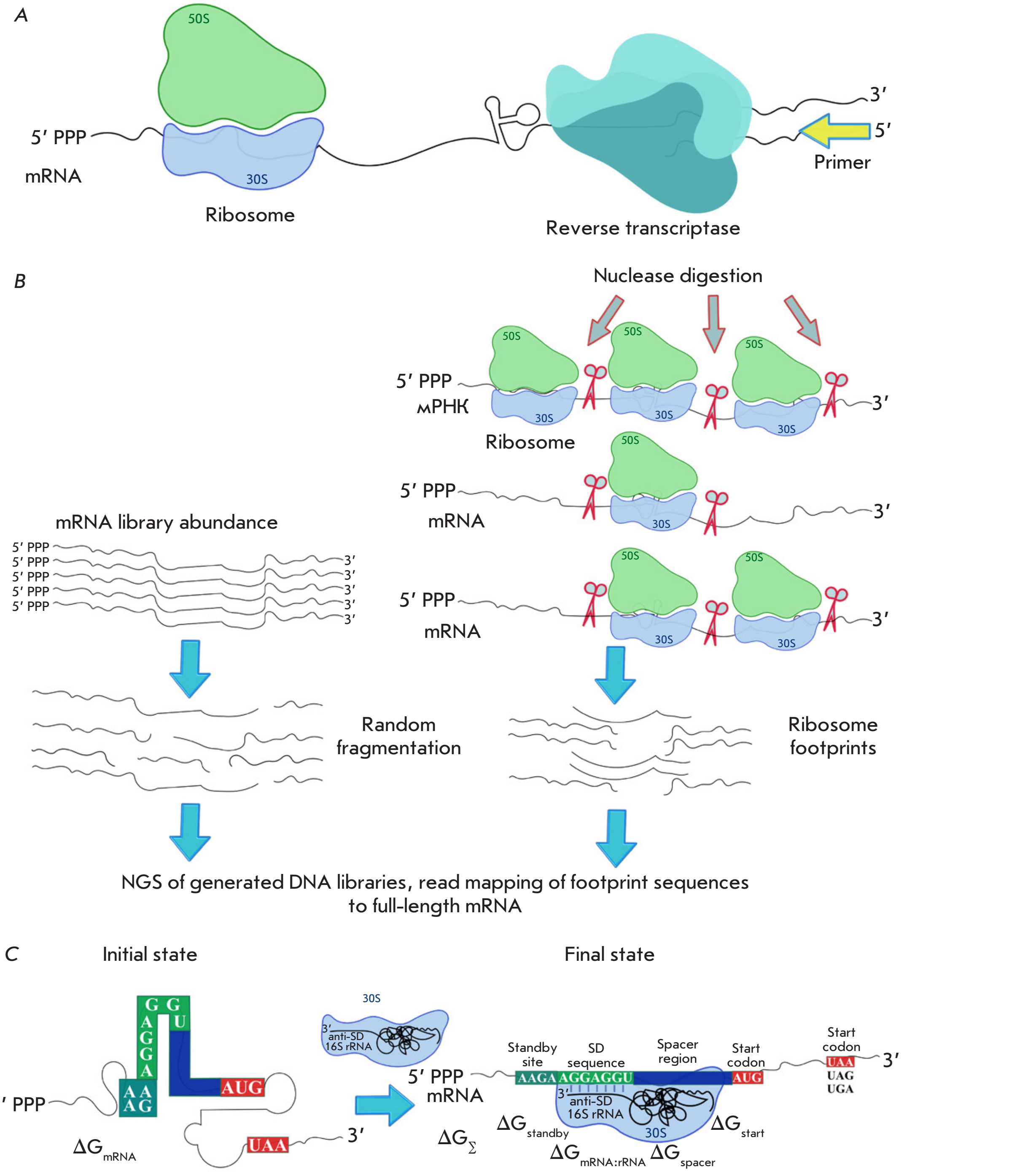


Fig. 2. (*A*) – The principle of the toeprinting technique. Stable ribosome complexes stop reverse transcriptase at a specific mRNA position, thus generating short cDNA products of a specific length. Primers for reverse transcriptase can be radioactively or fluorescently labeled. (*B*) – The scheme of the ribosome profiling method (Ribo-seq). After translation initiation, mRNA is cut by a specific nuclease at the sites where it is not protected by ribosomes. In parallel, the original mRNA library is prepared for sequencing by random fragmentation. It will be used as a reference sequence. All obtained ribosome footprints are used to prepare a DNA library, which is further deeply sequenced. Based on the NGS results, footprint sequence reads are mapped to full-length mRNA. (*C*) – The thermodynamic model of bacterial translation initiation. Changes in free energy during the initiation stage depend on the five types of molecular interactions defining the initial and the final states of the system. The drawing was executed in the Inkscape software

the SD motifs can be attributed to pauses at glycine codons and the impossibility of isolating the entire population of ribosome-protected mRNA fragments. A conclusion has been drawn that the SD motifs are probably not the main cause of the multiple pauses noted during translation *in vivo* [65].

The biophysical models allow one to assess the efficiency of biomolecule interactions, including the mRNA–ribosome ones. The thermodynamic model can be used as an example (*Fig. 2C*) [32]; this model simultaneously estimates the strength of the molecular interactions of the 30S complex with the mRNA transcript, calculates the Gibbs free energy for each element within a particular mRNA, and predicts the translation initiation rate: the higher energy needs to be spent to unfold mRNA elements, the lower the translation initiation rate is. The presented model can be used both to predict the relative translation initiation rate of an existing 5' UTR with the identified RBS and to design an RBS sequence ensuring the required translation initiation rate [18, 32].

The Flow-seq method used for a library of plasmids carrying the fluorescent protein genes (the first one acting as an internal control, and expression of the second varying depending on the impact of the sequences obtained by randomization of 30 nucleotides in the coding region of the gene immediately downstream of the start codon) allowed one to divide the resulting library (over 30×10^3 mRNA variants) according to translation efficiency [28]. Further analysis showed that the translation efficiency of mRNAs carrying the SD-like sequences was reduced, and that the proportion of such mRNAs in the set of efficiently translated mRNAs also declined, being indicative of the negative effect of the SD sequences in this mRNA region on the protein synthesis and, in turn, supporting earlier findings obtained for a limited set of model mRNAs [66].

Interestingly, the distribution of the binding energies of the anti-SD sequences among efficiently translated mRNAs is similar to that in natural *E. coli* genes. Moreover, individual constructs carrying the SD sequences in the sliding window of the initial coding region immediately downstream of the start codon and having similar energies of secondary structure folding have been designed, and their translation efficiency has been evaluated. Hence, the findings obtained are consistent with the results of the data analysis performed after using the Flow-seq method [28].

THE SCHEME OF THE FLOW-seq TECHNIQUE, THE FEATURES AND RESULTS OF ITS APPLICATION

Thousands of reporter constructs are often used to determine the effect of a certain factor or a set of factors on the expression level of a particular gene by sorting various promoter variants, 5' untranslated regions, and the individual sites in them (including the ribosome-binding sites (RBSs), the upstream regions (standby sites) or the downstream spacer sites), as well as the initial ramp regions of the coding sequence, either individually or simultaneously (*Table 1*). These plasmids typically carry two fluorescent protein genes: the first one acting as a sensor whose expression is sensitive to variable sites and the second one being used as an invariant internal control. The resulting sets of constructs are used to transform the bacterial strain suitable for further expression and sorting. Next, the fluorescence intensities of the two proteins in the cell pool are estimated using flow cytometry and cell groups/fractions characterized by approximately identical ratios of the measured fluorescence levels of these proteins are formed. Once the number of collected cells is increased, plasmids are isolated from the cells; the variable site is amplified and subjected to high-throughput sequencing in order to determine the DNA/RNA sequences in the particular fraction ensuring a particular level of reporter gene expression *(Fig. 3*).

This approach was applied to design a number of constructs simultaneously carrying different combinations of ribosome binding sites and promoters. The amounts of RNA and green fluorescent protein (GFP) synthesized by the cells transformed with each construct were compared to the amount of respective DNA, thus determining the transcription and translation efficiencies. The mCherry fluorescent protein gene, which was used as an internal control and carried a conserved promoter and ribosome binding site (RBS), was also inserted into the construct [67]. A set consisting of 12,653 plasmids with various combinations of 114 promoters and 111 RBS variants was eventually obtained. In order to estimate the steady-state DNA and RNA levels, deep sequencing of DNA (DNA-seq) and RNA (RNA-seq) from the cells in this phase was carried out. To assess the levels of the two fluorescent proteins, the cells were sorted according to the ratio between the GFP/mCherry fluorescence intensities. Plasmid DNA was isolated from cell populations with similar GFP/mCherry fluorescence intensity ratios and subjected to deep sequencing. The extracted sequences belonging to a particular group were tagged with group-specific barcode sequences, which were further used for searching for and sorting sequences into previously defined groups during the analysis of sequencing reads. The levels of two fluorescent proteins in the groups were then assessed; the GFP/mCherry ratio was defined as a measure of translation efficiency; the cells were subdivided into

Table 1. Application of the Flow-seq method to the analysis of the translation efficiency

mRNA elements	Number of variants in the generated libraries after Flow-seq	Variant types	Results	Reference
Promoters and ribosome binding sites (RBS) in 5’ UTR	11,894 (94%) out of 12,653 possible variants with combinations of 114 promoters and 111 RBSs (one combination resulted in the incompatible restriction site)	Taken from the available databases and generated using the RBS Calculator	The range of expression variations – four orders of magnitude. Promoter choice has the greatest effect on the RNA level and a smaller one on protein level, since its translation efficiency is also affected by the choice of the ribosome binding site and, potentially, other factors. 55% out of several hundreds of tested individual colonies were unmistakably identified during the Flow-seq analysis	[67]
Promoters and ribosome binding sites (RBS) in 5’ UTR	~ 500 combinations of 14 promoters and 22 RBSs for two detectable fluorescent proteins and more than 1,200 combinations from the randomized library	Specific variants and variants with randomized sequences in the elements under study	The dynamic range of expression – three orders of magnitude. The resulting combinations lead to the expression of a random gene (twofold variation in the expression level) with 93% reliability	[75]
Six nucleotides in the spacer region downstream of the SD sequence in the 5’ UTR and upstream of the start codon and the first six nucleotides following it (codons at positions +2 and +3 of the coding sequence (CDS)) …-SD-GAC-6N-AUG-$6N^{syn}$-…	13,914 (56%) variants for one protein and 25,861 (53%) variants for another protein out of 24,576 and 49,152 possible variants, respectively	Randomized spacer regions and codons at positions +2 and +3 with synonymous substitutions not changing the coding sequence of two sensor proteins	The range of expression variations – three orders of magnitude. The low GC-content and reduced stability of the secondary structure of the studied elements are important for the high expression level not limited by these determinants. The distribution of the protein fluorescence levels measured in several dozen colonies using a plate reader is consistent with the Flow-seq data	[71]
Four nucleotides in the spacer region downstream of the SD sequence in the 5’ UTR and upstream of the start codon …-SD-C-4N-CAU-AUG-…	249 (97%) out of the 256 possible variants	Randomized	The range of expression variations – two orders of magnitude. The predominant adenosine content and reduced cytidine content in efficiently translated variants. The low GC-content and reduced stability of the secondary structure of the studied elements are important for a high expression level. The SD-like sequences also occur only in the highly expressed variants	[39]
Six-nucleotide SD sequence in the 5’ UTR	4,066 (99%) out of the 4,096 possible variants	Randomized	The measured levels of proteins (fluorescent and five natural ones) for 91% of the sequence variants lay within the twofold range of variations in the expression level predicted using the EMOPEC tool that takes into account the context of the SD sequence, which minimized variations in the secondary structure	[76]
Standby sites of different lengths (20–164 nucleotides) upstream of the SD sequence, distal in the 5’ UTR	136 5’ UTRs with different lengths and secondary structures, shapes, and number of modules	Modeled variants	The range of variations in translation efficiency – two orders of magnitude. The rate of mRNA translation initiation is controlled by the surface area of single-stranded regions, partial unfolding of the RNA structure for minimizing the ribosome binding free energy penalty; there is no cooperative binding and, possibly, ribosome sliding in the analyzed region. The biophysical model for predicting the translation initiation rate has been developed and experimentally tested. The ribosome can easily bind to the modules of standby sites that are remote from the start codon and ensure high translation efficiency	[34]

Table 1 (continued)

mRNA elements	Number of variants in the generated libraries after Flow-seq	Variant types	Results	Reference
The ribosome binding site (RBS) in the 5' UTR with a fixed SD sequence (five nucleotides) and the variable standby site (four nucleotides) and the six-nucleotide spacer region RRRV-AGGAG-R-6N-AUG (R: A/G, V: A/G/C, N:A/U/C/G)	More than 20,000 (10%) out of ~ 200,000 possible variants for two fluorescent proteins	Randomized and partially specific positions with incomplete variations	The range of variations in translation efficiency – four orders of magnitude. The translation efficiency is significantly affected by conservation of the SD sequence, whereas the AC-rich spacer region is weakly dependent on the context. Low stability of the secondary structure of the studied region was observed for high expression. Replacement of the reporter protein with another one often had no effect on the overall trend in the distribution of the sequences defining a given protein synthesis level	[74]
Almost complete 5' UTR sequence (22 or 32 nucleotide long) GG-20N/30N-AUG…	11,692 (10^{-6}% out of the possible variants), 11,889 (10^{-12}%) for 20N and 30N, respectively; 48 natural variants with variations	Randomized, natural, and specific	The range of variations in translation efficiency – four orders of magnitude. Low stability of the secondary structure and conservation of the SD sequence in highly expressed variants were observed. The presence of AU-rich enhancers at the 5' terminus in the standby site, the low cytidine content, multiple SD sequences, and AG repeats in mRNA 5' UTRs ensure high translation efficiency in a number of cases	[2]
5' UTR sequences (2–60 nucleotides long) of the first genes of *E. coli* operons with GG at the 5' terminus retained during transcription GG-natural 5' UTR	648 (91%) out of the 713 possible variants 2–60 nucleotide long, (45%) out of all the 1,451 natural 5' UTRs of the first operon genes	Natural	The range of variations in translation efficiency – 30-fold. The RNA secondary structure and SD sequence affected the translation efficiency, but with lower variability compared to the randomized libraries. The low secondary structure stability and conservation of the SD sequence in highly expressed variants. The results of an estimation of the translation efficiency for individual 5' UTRs correlated with the ribosome profiling data	[77]
Sites in the promoter region, the standby site 10/20/30 nucleotides long, the 8-nucleotide spacer region 10N/20N/30N-SD-8N	~ 12,000 (a very small percentage of the possible variants)	Randomized	The range of variations in translation efficiency – five orders of magnitude. At a high expression level, low stability of the secondary structure of the studied region was observed	[72]
Promoters, ribosome binding sites (RBS), the first 13 amino acids of the protein-coding region	14,234 combinations of two promoters, four ribosome binding sites (RBSs), and sequences of N-terminal peptides corresponding to the first 13 amino acids in 137 natural *E. coli* genes	Natural	The range of variations in translation efficiency – more than two orders of magnitude. The use of rare codons at the N-terminus can increase expression 14-fold regardless of RBSs, ensuring a degree of translation efficiency. Reduction of secondary structure stability, rather than codon rarity itself, is responsible for increasing the expression level	[78]
The first six codons downstream of the start codon in the coding sequence	10	Natural	Reduction of secondary structure stability, rather than codon rarity itself, is responsible for increasing translation efficiency. Rare codons are often A/T-rich at position 3, which is more likely to correlate with increased expression than the synonymous G/C-ending codons	[81]
The first 10 codons downstream of the start codon in the coding sequence	More than 30,000	Randomized	Reduction of secondary structure stability, rather than codon rarity itself, is responsible for increasing translation efficiency. Codons located closer to the start codon have a significant effect on expression. Additional start codons in the reading frame facilitate translation. The presence of amino acids for the synthesis of which the cell expends a lot of resources, in the N-terminal motif of the protein negatively affected protein synthesis efficiency	[28]

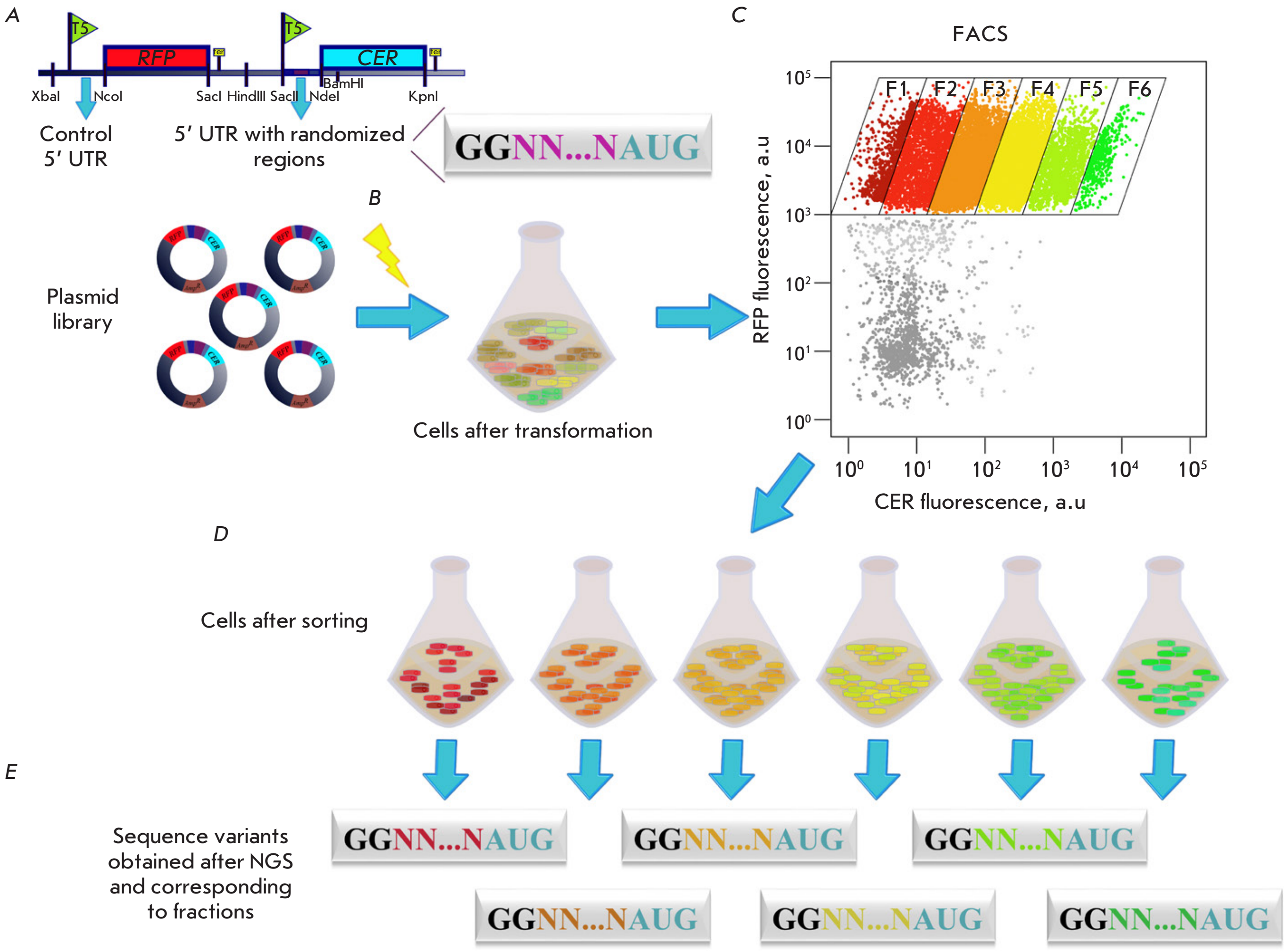


Fig. 3. The scheme of the Flow-seq method (as exemplified by working with randomized 5′ UTR upstream of the CER protein gene and control 5′ UTR upstream of the RFP protein gene). The stages of plasmid library construction, transformation, sorting, and sequencing are presented. (*A*) – Cloning of a randomized DNA fragment into a reporter vector upstream of the CER protein gene. A constant 5′ UTR is retained upstream of the RFP protein gene. (*B*) – Electroporation of the entire plasmid library into *E. coli* cells. (*C*) – Cell separation based on the CER/RFP fluorescence intensity ratio by a cell sorter. (*D*) – Cell fraction collection (e.g., F1–F6) according to the CER/RFP ratio. (*E*) – DNA isolation and randomized region amplification followed by high-throughput sequencing (NGS). The drawing was executed in the Inkscape software

three types according to this ratio: weak, medium and strong, and the corresponding sequences were identified. As anticipated, the cells in the library contained approximately identical levels of the mCherry protein, whose fluorescence intensities were characterized by the normal (Gaussian) distribution and varied within one order of magnitude, whereas the expression levels of the *gfp* gene varied by four orders of magnitude. A total of 282 individual colonies were verified by sequencing; 55% of these colonies were appropriate (i.e., contained error-free invariable sites, and the expected promoter variants and ribosome binding sites were identified for them without mutations). The fluorescence levels of most of these 55% appropriate promoter and RBS combinations were measured and subsequently used as a control set.

The results obtained by large-scale sequencing of DNA and RNA and the measured gene expression levels of the fluorescent proteins were used at the next stage as a platform for constructing the representative maps. When these maps were constructed, the transcription and translation levels were determined for each construct type with specific promoter and ribosome binding site variants (*Fig. 4*). Further

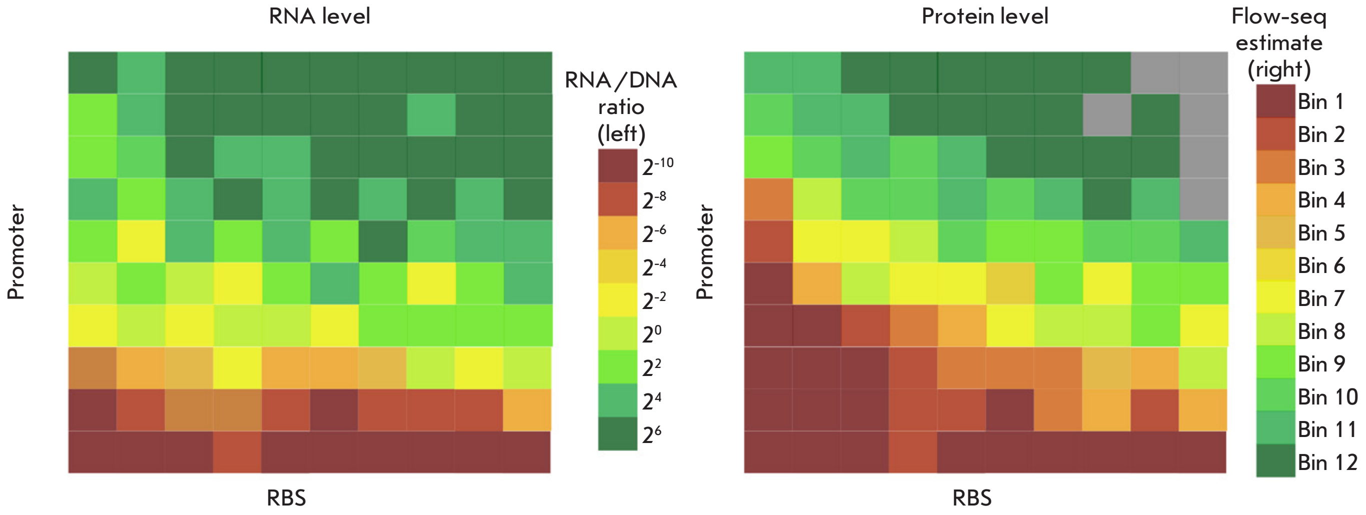


Fig. 4. A schematic image of the exemplary representative maps of RNA and protein synthesis efficiency levels. RNA (left) and protein (right) levels for a small set of constructs are gridded according to the identity of the promoters (the Y axis) and ribosome binding sites (RBS, the X axis). Promoters and RBSs are sorted in ascending order of the average efficiency of RNA and protein synthesis, respectively. Gray cells indicate constructs corresponding to levels below an empirically defined threshold. Scales of RNA levels (the RNA to DNA ratios) and protein levels (ratios of GFP (green) to RFP (red) fluorescence proteins) are shown to the right of their respective maps. The drawing was executed based on the source [67] in the Inkscape software

analysis allowed one to estimate the most efficient and inefficient combinations contained in the resulting construct library (*Table 2*) [67]. A comprehensive analysis of the variance (ANOVA) [68] of RNA and protein levels determined independently by both the promoter and the ribosome binding site was carried out. This approach also helped one to make allowance for the effects showing the association between the RNA level and the translation rate.

The programs written in R [69] and Python [70] and adapted to working with large datasets were used to visualize the resulting estimates. The ANOVA data made it possible to attribute the differences in RNA levels to the choice of promoter in 92.5% of cases, the choice of ribosome binding site in 3.8% of cases, while the remaining 3.7% of the differences could not be attributed to the choice of a variable element. The differences in the GFP protein levels were attributed to the promoter choice in 53.8% of the cases; the RBS choice, in 29.6% of cases; and the remaining percentage could be attributed to none of these two variable factors. Therefore, it was inferred that promoter choice had the greatest effect on the RNA level, while having a smaller impact on the protein level, since the translation efficiency is also affected by the choice of the ribosome binding site and, presumably, other factors as well [67].

A number of studies employing the Flow-seq method have investigated the effect of the sequences of 5’ untranslated regions of different lengths and their individual sites on the efficiency of the reporter fluorescent protein synthesis [2, 39, 71–74].

Variation in the spacer regions residing between the Shine–Dalgarno sequence and the start codon enabled the construction of small-sized libraries, where four and six nucleotides in a given site were randomly generated. A 100- [39] and 1,000-fold [71] difference between the highest and lowest produced protein level, respectively, was successfully obtained. In the former case, the most efficient and inefficient sequences included the following spacer sequences: cAAAAcau, cGAAAcau, cAUAAcau, cAUAUcau and cCCGCcau, cCUCUcau, cCGCUcau, cCCGUcau, respectively, by SD sequence (GAGG) flanking at the 5’ terminus and by the start codon (AUG) at the 3’ terminus. In the latter case, among the sequences residing downstream of the SD sequence (AAGAAGGA) and upstream of the start codon (AUG) and ensuring the highest expression, one can distinguish the gacUAGAGC, gacUGUAAG, gacAAAACC, and gacGUGGUU sequences. Interestingly, the CAAAAC sequence emerges as one of the most effective sequences in both cases.

In the former case, single-stranded oligonucleotides with four random nucleotides in the spacer region and the restriction sites required for subsequent insertion of the fluorescent protein CER gene into the vector upstream of the start codon were used for library generation. The resulting set of cells was

Table 2. Examples of the sequences of promoters and ribosome binding sites (RBS) ensuring inefficient and efficient expression

No.	Expression efficiency	Promoter	RBS
1	Inefficient expression	GGCGCGCCTCGACATTTATCCCTTGCGGCGAATACTTACAGCCATAGCAA	CACCATACACATATG
2		GGCGCGCCCTGATAGCTAGCTCAGTCCTAGGGATTATGCTAGCAGATG	ATCTTAATCTAGCGCGGGACAGTTTCATATG
3		GGCGCGCCTCGACAATTAATCATCCGGCTCGATACTTACAGCCATCGATT	TCTAGAGAAAGACCCGAGACACCATATG
4		GGCGCGCCCACGGTGTTAGACATTTATCCCTTGCGGCGAATACTTACAGCCATGTGAA	ATCTTAATCTAGCTTTGGAGTCTTTCATATG
5		GGCGCGCCTTGACAGCTAGCTCAGTCCTAGGGATTGTGCTAGCCAATC	TCTAGAGAAAGATTAGAGTCACCATATG
6		GGCGCGCCCACGGTGTTAGACAATTAATCATCCGGCTCGATACTTACAGCCATGATTC	ATCTTAATCTAGCCCGGGAGCATTTCATATG
7		GGCGCGCCTCGACATCAGGAAAATTTTTCTGATACTTACAGCCATGCGGA	TCTAGAGAAAGACAGGACCCACCATATG
8		GGCGCGCCCACGGTGTTAGACATCAGGAAAATTTTTCTGATACTTACAGCCATCGACC	TCTAGAGAAAGAGCCGACATACCATATG
9		GGCGCGCCTTTATAGCTAGCTCAGCCCTTGGTACAATGCTAGCGCCTG	ATCTTAATCTAGCCTGGGATCGTTTCATATG
10		GGCGCGCCTTTATGGCTAGCTCAGTCCTAGGTACAATGCTAGCCATAC	ATCTTAATCTAGCCCAGGAACGTTTCATATG
1	Efficient expression	GGCGCGCCTTGACATCGCATCTTTTTGTACCTATAATGTGTGGATAGAGT	AATCTCATATATCAAATATAGGGTGGATCATATG
2		GGCGCGCCAAAAAGAGTATTGACTTCAGGAAAATTTTTCTGTATAATGTGTGGATGTTCA	AATCTCATATATCAAATATAAGGCGGATCATATG
3		GGCGCGCCAAAAAGAGTATTGACTATTAATCATCCGGCTCGTATAATAGATTCATTGAAG	ATTAAAGAGGAGAAATTACATATG
4		GGCGCGCCTTGACATCGCATCTTTTTGTACCTATAATAGATTCATGATGA	AAAGATCTTTTAAGAAGGAGATATACATATG
5		GGCGCGCCTTGACATAAAGTCTAACCTATAGGATACTTACAGCCATACAAG	AAAGAGGAGAAATTACATATG
6		GGCGCGCCTTGACATCAGGAAAATTTTTCTGTAGATTTAACGTATAGGTA	AATCTCATAAATCAAATATAAGGGGGATCATATG
7		GGCGCGCCAAAAAGAGTATTGACTTCGCATCTTTTTGTACCTATAATAGATTCATTGCTA	GAATTCATTAAAGAGGAGAAAGGTCATATG
8		GGCGCGCCAAAAAGAGTATTGACTTCGCATCTTTTTGTACCCATAATTATTTCATTCACA	AATCTCATATCTCAAATATAAGGGGGATCATATG
9		GGCGCGCCAAAAAATTTATTTGCTTTTTATCCCTTGCGGCGATATAATAGATTCATCTTAG	AATCTCATAGATCAAATATAGGGGGGATCATATG
10		GGCGCGCCAAAAAATTTATTTGCTTTCGCATCTTTTTGTACCTATAATGTGTGGATAATAA	ATCTTAATCTAGCGGGGGAGAATTTCATATG

Note: examples of the combinations of promoter and ribosome binding site sequences were selected with allowance for the maximum and minimum RNA and translation levels, respectively, for efficient and inefficient protein expression; the sequences of restriction sites are underlined; the last five nucleotides in the promoter sequences act as the unique barcode for identification of the transcription initiation site. The sequences are shown in the 5′→3′ orientation.

subjected to sorting, and the selected variable plasmid regions were used for next-generation sequencing [39].

In the latter case, to optimize the synthesis of two specific proteins encoded by the *araH*^{WT} and *narK*^{WT} genes, their coding sequences were bound to the region encoding the TEV-GFP-His$_8$ additional sequence, where TEV is the recognition site of the tobacco etch virus protease (BTM/TEV); His$_8$ is a tag composed of eight His residues for further purification. Therefore, the measured GFP fluorescence can be indicative of the expression levels of the genes of interest. A vector comprising the aforedescribed complex coding region under the control of the T7 promoter, and two primers (the reverse one being invariant and the forward one containing six variable nucleotides upstream and downstream of the start codon; these nucleotides met the criteria of synonymous codon substitutions) was used for library construction. Expression was induced by IPTG; the cells were then sorted into separate fractions by FACS according to the intensity of the GFP protein fluorescence. Plasmid DNA libraries were then isolated from these fractions and subjected to high-throughput sequencing [71].

An analysis of the sequencing data for several tens of thousands of different mRNA variants obtained in the two experiments described above showed that the low GC-content and the absence (or minimization) of the mRNA secondary structure in the spacer region under study increased the amount of the synthesized protein [39, 71]. Therefore, it seems reasonable to use oligoadenylate or other A-rich spacers between the SD sequence and the start codon to increase the protein synthesis yield, while avoiding the use of cytidine bases, although one should not rule out certain specific mRNAs with A-rich spacer regions, which can mask the translation initiation site in their secondary structure if the beginning of the coding region is U-rich.

These results should be taken into account when designing reporter plasmids when there is a need for the expression levels of exogenous genes to be tuned according to specific biotechnological needs. For the coexpression of the genes whose products are supposed to be synthesized in a given stoichiometric ratio (e.g., when proteins are subunits of the heteromultimeric complex), the expression levels of these genes can be regulated by a judicious choice of the spacer regions.

Determining the sensitivity to minor variations in the sequence of the regulatory elements in the 5' UTR, such as the Shine–Dalgarno sequence, is rather challenging, since minor variations in the 5' UTR may lead to unpredictable changes in the gene expression level [34, 75]. The dependence of the translation efficiency on the 5' UTR sequence enables efficient and multiplex engineering, provided that the models being built can adequately predict these changes [73].

EMOPEC (Empirical Model and Oligos for Protein Expression Changes), another tool for predicting gene expression levels in bioengineering, has been developed; it is a nearly complete database of *gfp* expression levels measured using the Flow-seq method, depending on the presence of a particular SD sequence [76].

It is well known that the effect of a particular SD sequence largely depends on its genetic context [32]. Accordingly, special care should be taken when reapplying the measured expression levels in the bioengineering of metabolic pathways or synthetic biology, since the ribosome binding site depends in large part on the local secondary structure of mRNA. However, whereas the Shine–Dalgarno sequences can be modified by making minimal changes to the secondary structure in a given mRNA region, the relative order of expression level of a particular SD sequence will probably remain intact [73]. These features are taken into account when using the algorithm in the EMOPEC database, which allows one to test a wide range of gene expression levels, with minimal changes in the SD sequence. Therefore, parallel and efficient genome editing tuning gene expression levels becomes possible.

The Flow-seq method has been repeatedly used to gain insight into how the nucleotide sequences of different motifs of 5' UTRs affect the translation efficiency. In particular, the ribosome binding sites with a fixed SD sequence [74], 5' UTRs of different fixed lengths [2], or natural 5' UTRs of different lengths [77], as well as standby sites and spacer regions [72], were studied. An analysis of tens of thousands of tested variants showed that the variation in the efficiency of the reporter protein synthesis can amount to four, and even five, orders of magnitude. Moreover, replacement of one reporter protein with another one often did not affect the general trend of sequence distribution, which sets a particular level of protein biosynthesis, indicating that these changes are determined specifically by variable mRNA regions. Similar observations relating to the low stability of the secondary structure and the conservation of the SD sequence were made for the variants determining a high translation efficiency [2]. The same factors were found to be significant for the translation efficiency of the reporter gene preceded by a set of natural 5' UTRs; however, in this case, the variability of the translation efficiency was much lower than it was for the library of fully randomized 5' UTR sequences [77].

There were also individual cases being indicative of the presence of AU-rich enhancers at the 5' terminus at the standby site, low abundance of cytidine bases, multiple SD sequences, and AG repeats in the mRNA 5' UTRs, which provide the high reporter protein level [2].

A similar approach was also used to elucidate the effect of rare codons at the beginning of the mRNA coding region on the translation efficiency [78]. According to observations, rare codons are more frequently found at the beginning of the coding region of natural genes, especially the highly expressed ones, which may be important for ensuring the high protein synthesis level [64, 79–82]. According to other data, codon rarity at the beginning of the coding region is simply a consequence of a selection driven by the urge to minimize the secondary structure at the beginning of the mRNA coding region [19, 78, 82]. In the research literature, there is an ongoing discussion about the causes and consequences of rare codon clusters at the beginning of coding regions and how these clusters affect the translation efficiency. The potential reasons for the diverging opinions can lie in the collection peculiarities of the data on which these opinions are based. In particular, different research groups used natural [79–84] or synthetic sequences [80, 85–90], as well as slightly different methods of analysis [79–90], in drawing their conclusions.

In order to elucidate the reasons for the increased abundance of rare codons at the beginning of the coding region of bacterial genes and its functional role, a large library comprising 14,234 combinations of two promoters (strong and weak ones), four ribosome binding sites (strong, medium, weak, and natural ones), and sequences of the first 13 codons of 137 *E. coli* genes was constructed based on an oligonucleotide array. These regulatory elements were placed upstream of the gene encoding the super-folder green fluorescent protein (sfGFP) in the plasmid from which the mCherry protein is constitutively coexpressed [78]. The DNA, RNA, and protein levels were measured in the entire constructed library using DNA-seq, RNA-seq, and Flow-seq, respectively.

According to the "codon ramp" hypothesis, the first N-terminal codons in the coding region are slowly translated, which subsequently reduces ribosome stalling during protein synthesis [79, 88, 89]. The increase in the translation efficiency in the presence of rare codons at the beginning of the coding region can be attributed to changes in the mRNA secondary structure rather than to codon rarity [78]. Finally, the ribosome occupancy profiles have demonstrated that tRNA concentration, which actually is responsible for the efficiency of codon usage, does not correlate with the translation rate. Specific rare codons can create internal motifs similar to the SD sequence; in turn, they can affect the translation efficiency in *E. coli* cells [64]. Searching for an association between the internal SD-like motifs and variations in expression has revealed a weak but statistically significant relationship.

A study focusing on the effect of synonymous mutations on the translation efficiency has led to the following conclusion: the presence of rare codons in *E. coli*, often A/T-rich at position 3, is more likely to correlate with increased expression than the presence of synonymous G/C-ending codons, being indicative of an association with the mRNA secondary structure [85]. It has also been shown that reduction of GC content correlates with increased protein expression [78]. By predicting the RNA secondary structure for the first 120 bases of each transcript using the NUPACK software specializing in nucleic acid folding [91], it was found that the increase in strength of the secondary structure correlated with a reduction in the expression level, which explained why variation was more significant than any other change assessed previously [78].

More than 30 × 10^3 codon variants at positions 2–11 of the coding region of the reporter fluorescent protein obtained by randomization of the first 30 nucleotides downstream of the start codon were subsequently analyzed. The gene encoding the second fluorescent protein remained unchanged and was used as an internal control. The constructed plasmid library was examined using the Flow-seq method [28], making it possible to confirm that the mRNA secondary structure has a negative effect on the translation efficiency, while no positive role of the rare codons at the beginning of the coding region in gene expression was observed.

Meanwhile, the following patterns have been revealed. Some codons residing at the beginning of the coding region have a positive (AUG, AGA, GUA, GCA, CAC, CGA, UAC, AAA encoding additional Met along with the initiator one, the positively charged amino acids Arg, Lys, His, hydrophobic aliphatic Ala, Val and aromatic Tyr), while some others have a negative (CUC, CCC, CCG, CUG, GGA, GGG, GGC, GCC encoding hydrophobic aliphatic amino acids and amino acids with more or less conformational freedom compared to the rest of the amino acids Leu, Pro, Gly, and Ala) effect on the expression level. The closer the respective codon is to the initiator codon, the stronger the influence it has. Additional start codons in the reading frame facilitate translation. The presence of amino acids (the cell spends a lot of resources for synthesizing them) in the N-terminal motif of the protein negatively affects the synthesis efficiency of such proteins in a depleted environment.

Application of the Flow-seq method is not limited to the provided examples. This technique is also employed to evaluate (using reporter constructs as biosensors in various bacterial strains, including knockout ones [92]) the effects on the glycolytic processes, assess terminator sequences [93], identify the genes involved in the changes in a particular metabolic pathway (using biosensor constructs [94]), and solve other problems (e.g., study splicing) [95].

THE CONTRIBUTION OF THE FLOW-seq METHOD TO SYNTHETIC BIOLOGY

Synthetic biology is a recent scientific discipline that deals with designing and creating living organisms or individual processes occurring in natural organisms [96–98]. This discipline has emerged and has been developing through a combination of genetic engineering and recombinant DNA technologies with computational modeling. Therefore, synthetic biology seeks to identify the behavior of organisms and the processes occurring in them in order to subsequently modify and combine them to solve complex specific problems. For synthetic multicomponent systems to work reliably, the proteins comprising the system need to form at customized ratios [97].

Three calculator programs have been developed for estimating the translation efficiency based on the 5' UTR mRNA sequences, since the overall translation rate is believed to be proportional to the translation initiation rate. These calculators were shown to adequately estimate the protein synthesis level.

The RBS Calculator was the first one to appear among the three calculators [33, 99]. It relied upon the thermodynamic model studied previously and was a predictive design method for ensuring controlled translation initiation and protein synthesis in bacteria [32, 33]. This method allows one to vary the translation efficiency within the range of five orders of magnitude [33, 34]. However, the predictions made using the RBS Calculator are not always consistent with the experimental data obtained by Flow-seq or by testing individual reporter constructs [2].

The UTR Designer (or UTR Library Designer) is another computational method for modeling 5' UTR sequences capable of predicting the translation efficiency according to the mRNA sequence carrying a particular 5' UTR [100, 101]. Being similar to the RBS Calculator, this method employs a thermodynamic parameter defined as the difference in the Gibbs free energies before and after the assembly of the 30S translation initiation complex on mRNA and takes into account the affinity of ribosome interaction, as well as the availability of mRNA and ribosome. Like the RBS Calculator, this software has two engineering modes: in the forward-engineering mode, it generates a 5' UTR with a specified translation efficiency level of the target protein sequence. In the reverse-engineering mode, the calculator predicts the level of protein synthesis from the inserted mRNA sequence carrying the 5' UTR and the first 35 nucleotides of the protein-coding region. The operational principle of the described method of constructing the mRNA library with different 5' UTRs is to generate 5' UTR sequences by generating random nucleotide sequences and combinatorial enumeration of construction variants with a choice of those capable of providing the desired protein translation level. Moreover, there is a constant portion of the 5' UTR which must be present in the resulting sequence: in this case, the combinatorial enumeration will refer exclusively to its environment. This method was validated for two libraries of 5' UTRs carrying 16 sequences characterized by different translation levels lying in a given range using a fluorescent reporter; the *in silico* predictions agreed well with the *in vivo* data [100]. However, the predictions made using this approach are sometimes far from correlating with the *in vivo* results obtained for other 5' UTR sequence samples in the selected range of protein synthesis efficiencies.

Like the previous two calculators, the third one, RBS Designer, calculates the free energies but differs in the method used for predicting the translation rate. Relying on the steady-state kinetic model, this calculator estimates the probability of binding between a particular mRNA and the ribosome (translation efficiency), according to the chances for availability of the RBS-carrying mRNA region and affinity of ribosome binding. Each calculator is characterized by similar prediction accuracy [97].

Several prediction models have been reported thus far. They were constructed due to the vast amount of data obtained by large-scale sequencing, the analysis of various libraries, and the findings obtained using other genetic engineering techniques. A good example is the potential prediction of translation initiation sites, which is useful for localizing protein-coding gene sites during computer-assisted annotation of bacterial and archaeal genomes [102], and prediction of putative genomic sequences that correspond to functional RNA motifs [103], or prediction of gene expression levels with new combinations of genetic elements [75].

Even experimental verification of the translation efficiency determined by any binding site in a model system cannot guarantee that an identical efficiency will be achieved if the coding region sequence is replaced. Such is the case due to secondary structure formation when the coding region and the 5' UTR are complementary. A study using specially designed bicistronic constructs was conducted in order to increase the predictability of the expression level of any gene expressed in a heterologous system. In that study, a conventional short open reading frame was located upstream of the reporter coding region whose expression efficiency was measured by flow cytometry. The reading frames overlapped within the randomized translation re-initiation site. Therefore, it was found that re-initiation eliminates the dependence of the translation efficiency on the coding region of the second gene. Both *gfp* and *rfp* were used as the second gene in this synthetic operon. The resulting expression levels of these different genes correlated well with each other [75].

Hence, experimental determination of the expression efficiency by flow cytometry or Flow-seq can be directly and reliably employed for generating expression constructs in synthetic biology.

CONCLUSIONS

The Flow-seq technique combines flexible genetic bioengineering approaches and cell sorting based on flow cytometry and high-throughput sequencing of DNA to comprehensively assess genotype–phenotype associations. One of the applications of Flow-seq is in the study of the effect of specific regulatory elements on protein synthesis (*Table 1*). Designing tailored changes based on reporter constructs using the fluorescent protein genes allows one to quickly and efficiently elucidate the contribution of specific variants of regulatory sequences to the protein synthesis efficiency. Like other methods used to study the effect of 5' untranslated region elements in mRNA on the translation efficiency, this approach has its own peculiarities that should be taken into account when planning a complex multi-step experiment. Although the method discussed in this review has great potential, its application has some limitations, primarily caused by the challenges arising at different stages, such as DNA library cloning, sorting of cells with different ratios of fluorescence intensities of the reporter proteins, high-throughput sequencing, analysis of the reads obtained, and further calculations. Another limitation is that only two fluorescent proteins or other detectable reagents of such type are used, since there is a risk of fluorescence spectral overlapping for these proteins and, therefore, signal registration errors. Nonetheless, the Flow-seq method is widely used in various research fields and has remained relevant for many years. ●

This work was supported by the Russian Foundation for Basic Research (grants Nos. 17-00-00369, 17-00-00366, and 17-00-00367).

REFERENCES

1. Saier M.H. Jr. // J. Bacteriol. 2019. V. 201. № 15. P. e00091–e119.
2. Evfratov S.A., Osterman I.A., Komarova E.S., Pogorelskaya A.M., Rubtsova M.P., Zatsepin T.S., Semashko T.A., Kostryukova E.S., Mironov A.A., Burnaev E., et al. // Nucl. Acids Res. 2017. V. 45. № 6. P. 3487–3502.
3. Brenneis M., Soppa J. // PLoS One. 2009. V. 4. № 2. P. e4484.
4. Shine J., Dalgarno L. // Nature. 1975. V. 254. P. 34–38.
5. Shine J., Dalgarno L. // Proc. Natl. Acad. Sci. USA. 1974. V. 71. № 4. P. 1342–1346.
6. Kozak M. // Gene. 2005. V. 361. P. 13–37.
7. Shultzaberger R.K., Bucheimer R.E., Rudd K.E., Schneider T.D. // J. Mol. Biol. 2001. V. 313. № 1. P. 215–228.
8. Rudd K.E. // Nucl. Acids Res. 2000. V. 28. № 1. P. 60–64.
9. Ma J., Campbell A., Karlin S. // J. Bacteriol. 2002. V. 184. P. 5733–5745.
10. Gardner P.P., Eldai H. // Nucl. Acids Res. 2015. V. 43. № 2. P. 691–698.
11. Schluenzen F., Tocilj A., Zarivach R., Harms J., Gluehmann M., Janell D., Bashan A., Bartels H., Agmon I., Franceschi F., et al. // Cell. 2000. V. 102. P. 615–623.
12. Kaminishi T., Wilson D.N., Takemoto C., Harms J.M., Kawazoe M., Schluenzen F., Hanawa-Suetsugu K., Shirouzu M., Fucini P., Yokoyama S. // Structure. 2007. V. 15. P. 289–297.
13. Arenz S., Wilson D.N. // Cold Spring Harb. Perspect. Med. 2016. V. 6. № 9. P. a025361.
14. Wegmann U., Horn N., Carding S.R. // Appl. Environ. Microbiol. 2013. V. 79. № 6. P. 1980–1989.
15. Nakagawaa S., Niimurab Y., Miurac K.-i., Gojobori T. // Proc. Natl. Acad. Sci. USA. 2010. V. 107. № 14. P. 6382–6387.
16. Vimberg V., Tats A., Remm M., Tenson T. // BMC Mol. Biol. 2007. V. 8. P. 100.
17. Osterman I.A., Evfratov S.A., Sergiev P.V., Dontsova O.A. // Nucl. Acids Res. 2013. V. 41. P. 474–486.
18. Chen H., Bjerknes M., Kumar R., Jay E. // Nucl. Acids Res. 1994. V. 22. P. 4953–4957.
19. Gu W., Zhou T., Wilke C.O. // PLoS Comput. Biol. 2010. V. 6. P. e1000664.
20. Gingold H., Pilpel Y. // Mol. Systems Biol. 2011. V. 7. P. 481.
21. de Smit M.H, van Duin J. // Proc. Natl. Acad. Sci. USA. 1990. V. 87. P. 7668–7672.
22. Sean M.S., Simpson J. // Mol. Cell. 2006. V. 22. P. 105–115.
23. Ban N., Beckmann R., Cate J.H., Dinman J.D., Dragon F., Ellis S.R., Lafontaine D.L., Lindahl L., Liljas A., Lipton J.M., et al. // Curr. Opin. Struct. Biol. 2014. V. 24. P. 165–169.
24. Laursen B.S., Sorensen H.P., Mortensen K.K., Sperling-Petersen H.U. // Microbiol. Mol. Biol. Rev. 2005. V. 69. P. 101–123.

25. Lauber M.A., Rappsilber J., Reilly J.P. // Mol. Cell. Proteomics. 2012. V. 11. P. 1965–1976.
26. Stenström C.M., Isaksson L.A. // Gene. 2002. V. 288. P. 1–8.
27. Gonzalez de Valdivia E.I., Isaksson L.A. // Nucl. Acids Res. 2004. V. 32. № 17. P. 5198–5205.
28. Osterman I.A., Chervontseva Z.S., Evfratov S.A., Sorokina A.V., Rodin V.A., Rubtsova M.P., Komarova E.S., Zatsepin T.S., Kabilov M.R., Bogdanov A.A., et al. // Nucl. Acids Res. 2020. V. 48. P. 6931–6942.
29. Park Y.S., Seo S.W., Hwang S., Chu H.S., Ahn J.-H., Kim T.-W., Kim D.-M., Jung G.Y. // Biochem. Biophys. Res. Commun. 2007. V. 356. № 1. P. 136–141.
30. Barendt P.A., Shah N.A., Barendt G.A., Sarkar C.A. // PLoS Genet. 2012. V. 8. P. e1002598.
31. Barendt P.A., Shah N.A., Barendt G.A., Kothari P.A., Sarkar C.A. // ACS Chem. Biol. 2013. V. 8. № 5. P. 958–966.
32. Salis H.M., Mirsky E.A., Voigt C.A. // Nat. Biotechnol. 2009. V. 27. № 10. P. 946–950.
33. Salis H.M. // Meth. Enzymol. 2011. V. 498. P. 19–42.
34. Borujeni A.E., Channarasappa A.S., Salis H.M. // Nucl. Acids Res. 2014. V. 42. № 4. P. 2646–2659.
35. Hofacker I.L. // Nucl. Acids Res. 2003. V. 31. № 13. P. 3429–3431.
36. Farasat I., Kushwaha M., Collens J., Easterbrook M., Guido M., Salis H.M. // Mol. Syst. Biol. 2014. V. 10. P. 731.
37. Nakeff A., Valeriote F., Gray J.W., Grabske R.J. // Blood.1979. V. 53. № 4. P. 732–745.
38. Solieri L., Dakal T.C., Giudici P. // Ann. Microbiol. 2012. V. 63. P. 21–37.
39. Komarova E.S., Chervontseva Z.S., Osterman I.A., Evfratov S.A., Rubtsova M.P., Zatsepin T.S., Semashko T.A., Kostryukova E.S., Bogdanov A.A., Gelfand M.S., et al. // Microb. Biotechnol. 2020. V. 13. P. 1254–1261.
40. Kim D., Hong J.S.-J., Qiu Y., Nagarajan H., Seo J.-H., Cho B.K., Tsai S.F., Palsson B.Ø. // PLoS Genet. 2012. V. 8. № 8. P. e1002867.
41. Lesnik E.A., Fogel G.B., Weekes D., Henderson T.J., Levene H.B., Sampath R., Ecker D.J. // BioSystems. 2005. V. 80. P. 145–154.
42. Gould P.S., Bird H., Easton A.J. // BioTechniques. 2005. V. 38. P. 397–400.
43. Shirokikh N.E., Alkalaeva E.Z., Vassilenko K.S., Afonina Z.A., Alekhina O.M., Kisselev L.L., Spirin A.S. // Nucl. Acids Res. 2010. V. 38. № 3. P. e15.
44. Wen J.-D., Kuo S.-T., Chou H.-H.D. // RNA Biol. 2021. V. 18. № 11. P. 1489–1500.
45. Tzareva N.V., Makhno V.I., Boni I.V. // FEBS Lett. 1994. V. 337. P. 189–194.
46. Zheng X., Hu G.Q., She Z.S., Zhu H. // BMC Genomics. 2011. V. 12. P. 361.
47. Ingolia N.T., Ghaemmaghami S., Newman J.R., Weissman J.S. // Science. 2009. V. 324. № 5924. P. 218–223.
48. Andreev D.E., O'Connor P.B., Fahey C., Kenny E.M., Terenin I.M., Dmitriev S.E., Cormican P., Morris D.W., Shatsky I.N., Baranov P.V. // Elife. 2015. V. 4. P. e03971.
49. Andreev D.E., O'Connor P.B., Zhdanov A.V., Dmitriev R.I., Shatsky I.N., Papkovsky D.B., Baranov P.V. // Genome Biol. 2015. V. 16. № 1. P. 90.
50. Meydan S., Marks J., Klepacki D., Sharma V., Baranov P.V., Firth A.E., Margus T., Kefi A., Vázquez-Laslop N., Mankin A.S. // Mol. Cell. 2019. V. 74. № 3. P. 481–493.e6.
51. Brar G.A., Weissman J.S. // Nat. Rev. Mol. Cell. Biol. 2015. V. 16. № 11. P. 651–664.
52. Reid D.W., Shenolikar S., Nicchitta C.V. // Methods. 2015. V. 91. P. 69–74.
53. Ingolia N.T., Hussmann J.A., Weissman J.S. // Cold Spring Harb. Perspect. Biol. 2019. V. 11. № 5. P. a032698.
54. Weaver J., Mohammad F., Buskirk A.R., Storz G. // mBio. 2019. V. 10. № 2. P. e02819–18.
55. Meydan S., Klepacki D., Mankin A.S., Vázquez-Laslop N. // Meth. Mol. Biol. 2021. V. 2252. P. 27–55.
56. Vazquez-Laslop N., Sharma C.M., Mankin A., Buskirk A.R. // J. Bacteriol. 2022. V. 204. № 1. P. e0029421.
57. O'Connor P.B., Andreev D.E., Baranov P.V. // Nat. Commun. 2016. V. 7. P. 12915.
58. Andreev D.E., O'Connor P.B., Loughran G., Dmitriev S.E., Baranov P.V., Shatsky I.N. // Nucl. Acids Res. 2017. V. 45. № 2. P. 513–526.
59. Glaub A., Huptas C., Neuhaus K., Ardern Z. // J. Biol. Chem. 2020. V. 295. № 27. P. 8999–9011.
60. Gerashchenko M.V., Gladyshev V.N. // Nucl. Acids Res. 2017. V. 45. № 2. P. e6.
61. Marks J., Kannan K., Roncase E.J., Klepacki D., Kefi A., Orelle C., Vázquez-Laslop N., Mankin A.S. // Proc. Natl. Acad. Sci. USA. 2016. V. 113. № 43. P. 12150–12155.
62. Vázquez-Laslop N., Mankin A.S. // Annu. Rev. Microbiol. 2018. V. 72. P. 185–207.
63. Svetlov M.S., Koller T.O., Meydan S., Shankar V., Klepacki D., Polacek N., Guydosh N.R., Vázquez-Laslop N., Wilson D.N., Mankin A.S. // Nat. Commun. 2021. V. 12. № 1. P. 2803.
64. Li G.W., Oh E., Weissman J.S. // Nature. 2012. V. 484. № 7395. P. 538–541.
65. Mohammad F., Woolstenhulme C.J., Green R., Buskirk A.R. // Cell Rep. 2016. V. 14. № 4. P. 686–694.
66. Jin H., Zhao Q., Gonzalez de Valdivia E.I., Ardell D.H., Stenström M., Isaksson L.A. // Mol. Microbiol. 2006. V. 60. № 2. P. 480–492.
67. Kosuri S., Goodman D.B., Cambray G., Mutalik V.K., Gao Y., Arkin A.P., Endy D., Church G.M. // Proc. Natl. Acad. Sci. USA. 2013. V. 110. № 34. P. 14024–14029.
68. Mutalik V.K., Guimaraes J.C., Cambray G., Mai Q.A., Christoffersen M.J., Martin L., Yu A., Lam C., Rodriguez C., Bennett G., Keasling J.D., Endy D., Arkin A.P.// Nat. Methods. 2013. V. 10. P. 347–353.
69. Dessau R.B., Pipper C.B. // Ugeskr. Laeger 2008. V. 170. P. 328–330.
70. Sanner M.F. // J. Mol. Graph. Model. 1999. V. 17. P. 57–61.
71. Mirzadeh K., Martinez V., Toddo S., Guntur S., Herrgard M.J., Elofsson A., Norholm M.H., Daley D.O. // ACS Synth. Biol. 2015. V. 4. P. 959–965.
72. Sauer C., van Themaat E.V.L., Boender L.G.M., Groothuis D., Cruz R., Hamoen L.W., Harwood C.R., van Rij T. // ACS Synth. Biol. 2018. V. 7. № 7. P. 1773–1784.
73. Klausen M.S., Sommer M.O.A. // Meth. Mol. Biol. 2018. V. 1671. P. 3–14.
74. Duan Y., Zhang X., Zhai W., Zhang J., Zhang X., Xu G., Li H., Deng Z., Shi J., Xu Z. // ACS Synth. Biol. 2022. V. 11. № 8. P. 2726–2740.
75. Mutalik V.K., Guimaraes J.C., Cambray G., Lam C., Christoffersen M.J., Mai Q.-A., Tran A.B., Paull M., Keasling J.D., Arkin A.P., et al. // Nat. Methods. 2013. V. 10. P. 354–360.
76. Bonde M.T., Pedersen M., Klausen M.S., Jensen S.I., Wulff T., Harrison S., Nielsen A.T., Herrgård M.J., Sommer M.O. // Nat. Meth. 2016. V. 13. P. 233–236.
77. Komarova E.S., Slesarchuk A.N., Rubtsova M.P., Os-

terman I.A., Tupikin A.E., Pyshnyi D.V., Dontsova O.A., Kabilov M.R., Sergiev P.V. // Int. J. Mol. Sci. 2022. V. 23. № 20. P. 12293.

78. Goodman D.B., Church G.M., Kosuri S. // Science. 2013. V. 342. № 6157. P. 475–479.

79. Tuller T., Carmi A., Vestsigian K., Navon S., Dorfan Y., Zaborske J., Pan T., Dahan O., Furman I., Pilpel Y. // Cell. 2010. V. 141. № 2. P. 344–354.

80. Allert M., Cox J.C., Hellinga H.W. // J. Mol. Biol. 2010. V. 402. № 5. P. 905–918.

81. Pechmann S., Frydman J. // Nat. Struct. Mol. Biol. 2013. V. 20. № 2. P. 237–243.

82. Bentele K., Saffert P., Rauscher R., Ignatova Z., Blüthgen N. // Mol. Syst. Biol. 2013. V. 9. P. 675.

83. dos Reis M., Savva R., Wernisch L. // Nucl. Acids Res. 2004. V. 32. № 17. P. 5036–5044.

84. Shah P., Ding Y., Niemczyk M., Kudla G., Plotkin J.B. // Cell. 2013. V. 153. № 7. P. 1589–1601.

85. Kudla G., Murray A.W., Tollervey D., Plotkin J.B. // Science. 2009. V. 324. № 5924. P. 255–258.

86. Welch M., Govindarajan S., Ness J.E., Villalobos A., Gurney A., Minshull J., Gustafsson C. // PLoS One 2009. V. 4. № 9. P. e7002.

87. Zhou M., Guo J., Cha J., Chae M., Chen S., Barral J.M., Sachs M.S., Liu Y. // Nature. 2013. V. 495. № 7439. P. 111–115.

88. Navon S., Pilpel Y. // Genome Biol. 2011. V. 12. № 2. P. R12.

89. Tuller T., Waldman Y.Y., Kupiec M., Ruppin E. // Proc. Natl. Acad. Sci. USA. 2010. V. 107. № 8. P. 3645–3650.

90. Subramaniam A.R., Pan T., Cluzel P. // Proc. Natl. Acad. Sci. USA. 2013. V. 110. № 6. P. 2419–2424.

91. Zadeh J.N., Steenberg C.D., Bois J.S., Wolfe B.R., Pierce M.B., Khan A.R., Dirks R.M., Pierce N.A. // J. Comput. Chem. 2010. V. 32. № 1. P. 170–173.

92. Lehning C.E., Siedler S., Ellabaan M.M.H., Sommer M.O.A. // Metab. Eng. 2017. V. 42. P. 194–202.

93. Zhai W., Duan Y., Zhang X., Xu G., Li H., Shi J., Xu Z., Zhang X. // Synth. Syst. Biotechnol. 2022. V. 7. № 4. P. 1046–1055.

94. Glanville D.G., Mullineaux-Sanders C., Corcoran C.J., Burger B.T., Imam S., Donohue T.J., Ulijasz A.T. // mSystems. 2021. V. 6. № 1. P. e00933–20.

95. Cheung R., Insigne K.D., Yao D., Burghard C.P., Wang J., Hsiao Y.E., Jones E.M., Goodman D.B., Xiao X., Kosuri S. // Mol. Cell. 2019. V. 73. № 1. P. 183–194.e8.

96. Andrianantoandro E., Basu S., Karig D.K., Weiss R. // Mol. Systems Biol. 2006. V. 2. P. 2006.0028.

97. Reeve B., Hargest T., Gilbert C., Ellis T. // Front. Bioeng. Biotechnol. 2014. V. 2. P. 1–6.

98. Chappell J., Jensen K., Freemont P.S. // Nucl, Acids Res. 2013. V. 41. № 5. P. 3471–3481.

99. Zhang L., Lin X., Wang T., Guo W., Lu Y. // Bioresour. Bioprocess. 2021. V. 8. № 1. P. 58.

100. Seo S.W., Yang J.S., Kim I., Yang J., Min B.E., Kim S., Jung G.Y. // Metab. Eng. 2013. V. 15. P. 67–74.

101. Seo S.W., Yang J.S., Cho H.S., Yang J., Kim S.C., Park J.M., Kim S., Jung G.Y. // Sci. Rep. 2015. V. 4. № 1. P. 4515.

102. Zhu H., Wang Q. // Curr. Bioinformat. 2014. V. 9. P. 155–165.

103. Laserson U., Gan H.H., Schlick T. // Nucl. Acids Res. 2005. V. 33. № 18. P. 6057–6069.

Bulky Adducts in Clustered DNA Lesions: Causes of Resistance to the NER System

N. V. Naumenko[a], I. O. Petruseva[a], O. I. Lavrik[*]
Institute of Chemical Biology and Fundamental Medicine, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, 630090 Russia
[a] The authors have an equal contribution to the writing of the article.
[*]E-mail: lavrik@niboch.nsc.ru

Received May 31, 2022; in final form, October 18, 2022
DOI: 10.32607/actanaturae.11741

ABSTRACT **The nucleotide excision repair (NER) system removes a wide range of bulky DNA lesions that cause significant distortions of the regular double helix structure. These lesions, mainly bulky covalent DNA adducts, are induced by ultraviolet and ionizing radiation or the interaction between exogenous/endogenous chemically active substances and nitrogenous DNA bases. As the number of DNA lesions increases, e.g., due to intensive chemotherapy and combination therapy of various diseases or DNA repair impairment, clustered lesions containing bulky adducts may occur. Clustered lesions are two or more lesions located within one or two turns of the DNA helix. Despite the fact that repair of single DNA lesions by the NER system in eukaryotic cells has been studied quite thoroughly, the repair mechanism of these lesions in clusters remains obscure. Identification of the structural features of the DNA regions containing irreparable clustered lesions is of considerable interest, in particular due to a relationship between the efficiency of some antitumor drugs and the activity of cellular repair systems. In this review, we analyzed data on the induction of clustered lesions containing bulky adducts, the potential biological significance of these lesions, and methods for quantification of DNA lesions and considered the causes for the inhibition of NER-catalyzed excision of clustered bulky lesions.**

KEYWORDS **nucleotide excision repair, bulky DNA lesions, clustered DNA lesions.**

ABBREVIATIONS **AP site – apurinic/apyrimidinic site; B[a]P-dG – benzo[a]pyrene-guanine adduct; BER – base excision repair; BHD – β-hairpin domain of the XPC protein; CPD – cyclobutane pyrimidine dimer; ICL – interstrand DNA crosslink; IR – ionizing radiation; nAnt – non-nucleotide fragment of a DNA strand containing a bulky anthracenylcarbamoyl residue; nFlu – non-nucleotide fragment of a DNA strand containing a bulky fluorescein residue; NER – nucleotide excision repair.**

INTRODUCTION

The nucleotide excision repair (NER) system eliminates various DNA lesions, most of which are bulky adducts that introduce significant distortions into the regular double-stranded DNA structure. NER can be initiated via two pathways: the global genome (GG-NER) and transcription-coupled (TC-NER) ones. The transcription-coupled pathway recognizes lesions in the transcribed strands of active genes [1, 2]. TC-NER is triggered by stalling of the RNA polymerase II complex when the enzyme encounters a bulky lesion in the transcribed DNA strand. The GG-NER pathway removes lesions throughout the genome, including its non-transcribed regions and silent chromatin. In GG-NER, XPC factor complexes act as damage sensors. Starting from the second step of repair (damage verification), GG-NER and TC-NER involve the same set of protein factors and enzymes. DNA lesions are eliminated together with a 24–32-bp surrounding region. The resulting gap is filled by repair synthesis (*Fig. 1*) [3, 4].

Totally, NER involves more than 30 enzymes and protein factors that successively form in the DNA damage area variable on composition and structure complexes, which interact with DNA over two or three of the helix turns.

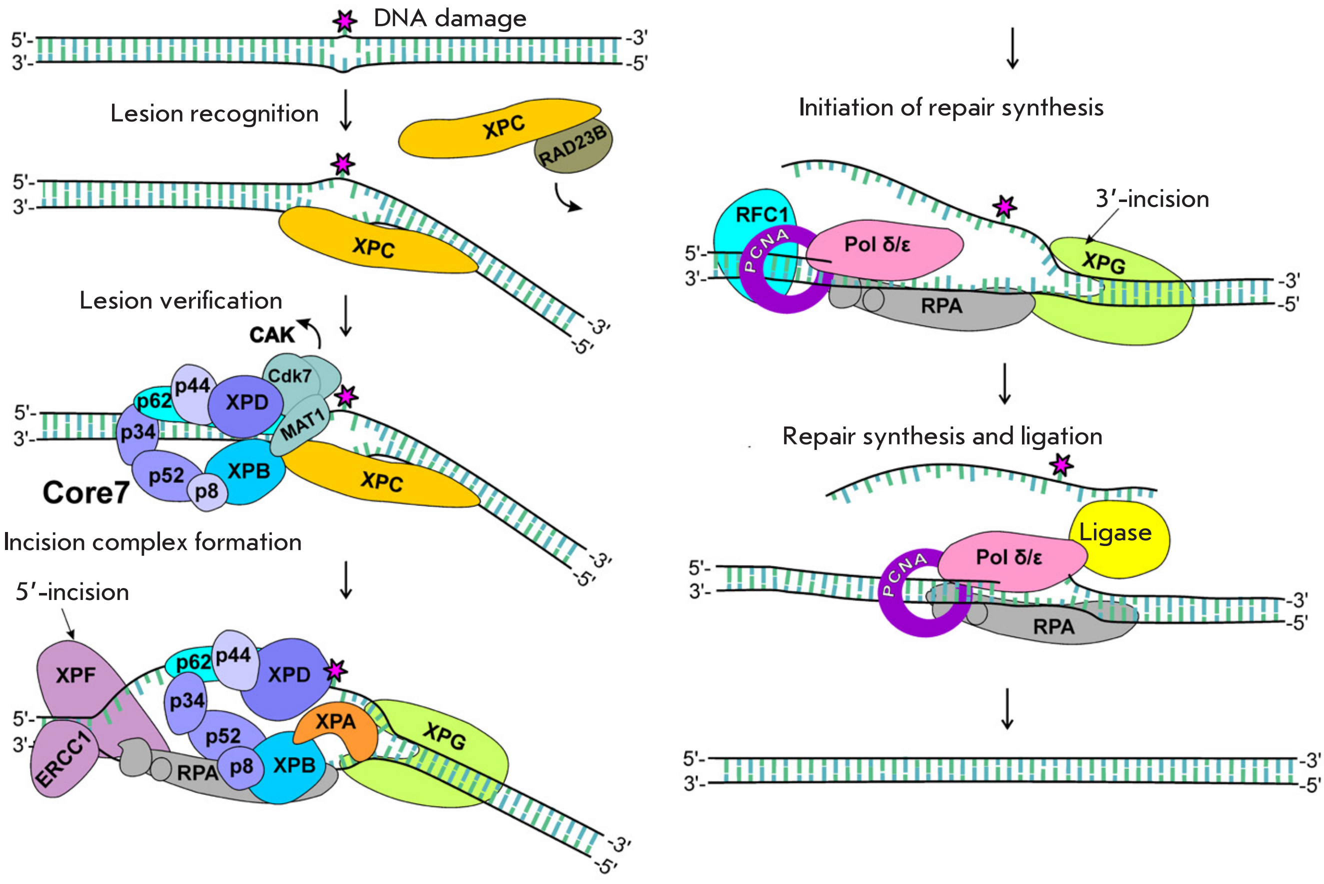


Fig. 1. Scheme of the global genome nucleotide excision repair

Several lesions within one or two DNA helical turns are called a clustered lesion (cluster) [5]. Clusters include various lesions: oxidized nitrogenous bases, AP sites, other non-bulky lesions, DNA strand breaks, and DNA fragments containing bulky adducts [5–7]. In recent years, great progress has been made in understanding NER repair of single lesions [8]. In contrast, the mechanism for the removal of clustered bulky lesions is much less studied. A number of studies have shown that the formation of an additional DNA lesion near a bulky adduct often reduces the efficiency of its removal by the NER system [9–11]. In addition, simultaneous excision of lesions in the opposite DNA strands may lead to the formation of double-strand breaks that are potentially lethal for the cell [12]. On the other hand, high activity of repair systems towards induced DNA lesions in tumor cells reduces the efficiency of antitumor drugs [13, 14]. Therefore, exploration of the mechanisms of interaction between repair proteins and clustered lesions and elucidation of any relationship between the structure of therapy-induced DNA lesions and their resistance to repair is of practical importance.

In this review, we analyzed data on the formation of clustered lesions containing bulky adducts and the potential biological significance of these lesions, considered inhibition of excision of bulky DNA lesions due to NER's unproductive binding of the XPC factor to damaged DNA, and addressed the structural features of the DNA regions containing clustered lesions resistant to NER.

THE ORIGIN AND TYPES OF NER-REPAIRABLE DNA LESIONS

Bulky DNA lesions, mainly covalent base adducts (*Fig. 2*), are induced by exposure to ultraviolet radiation (pyrimidine-(6,4)-pyrimidine photoproducts and cyclobutane pyrimidine dimers (lesion structures are shown in *Fig. 2A*) and strong ionizing radiation (IR) (e.g., oxidized 8,5'-cyclo-2'-deoxypurines, *Fig. 2B*, left; adducts of oxidized estrogen metabolites, *Fig. 2B*, right) [15–17]. Bulky DNA lesions are also induced by chemically active or cellular metabolism-activated substances: incomplete fuel combustion products (e.g., benzo[a]pyrene derivatives, *Fig. 2C*, left), tobacco smoke components (tobacco-specific nitrosamines, *Fig. 2C*, right [18–20]), DNA–protein crosslink-inducing agents [21], and some natural substances (e.g., aristolochic acids) [22]. Many of these lesions are difficult to repair and tend to accumulate in the body [10, 23, 24].

The cytostatic effect of many chemotherapeutic drugs is based on their ability to form bulky adducts upon interaction with DNA. These drugs include Pt-containing drugs (carboxyplatin, oxaliplatin, cisplatin; the structure of interstrand cisplatin crosslinked DNA is shown in *Fig. 2D*, top) [14, 25], alkylating nitrogen mustards (mechlorethamine, cyclophosphamide, and acylfulvene) [25, 26], minor groove ligands, mitomycins [27], and anthracycline drugs capable of forming covalent adducts with DNA in the presence of endogenous formaldehyde (*Fig. 2D*, bottom) [28].

METHODS FOR QUANTIFICATION OF BULKY LESIONS

Quantification of DNA lesions is a challenge, because the content of damaged nucleotides in total DNA is relatively small, and their structure and properties are diverse. A wide range of methods are used to detect and quantify bulky DNA adducts. Apart from the well-known single-cell electrophoresis under alkaline conditions (alkaline DNA-comet assay) [29], there are methods based on radioactive labeling, which are characterized by limited specificity but high sensitivity to detect one adduct per 10^9–10^{10} nucleotides [30–32]. In addition, there are more selective techniques based on the use of lesion-specific antibodies (the detection threshold is one adduct per 10^8 nucleotides) [18, 33, 34] and new variants of the polymerase chain reaction [35]. Quantification of lesions by atomic absorption spectrometry requires a 10–50 μL sample with an expected analyte concentration of 10^{-3} to 10^{-6} M [36].

Mass spectrometric techniques provide the highest quantification accuracy and specificity for lesions. The only limitation of mass spectrometry is that ac-

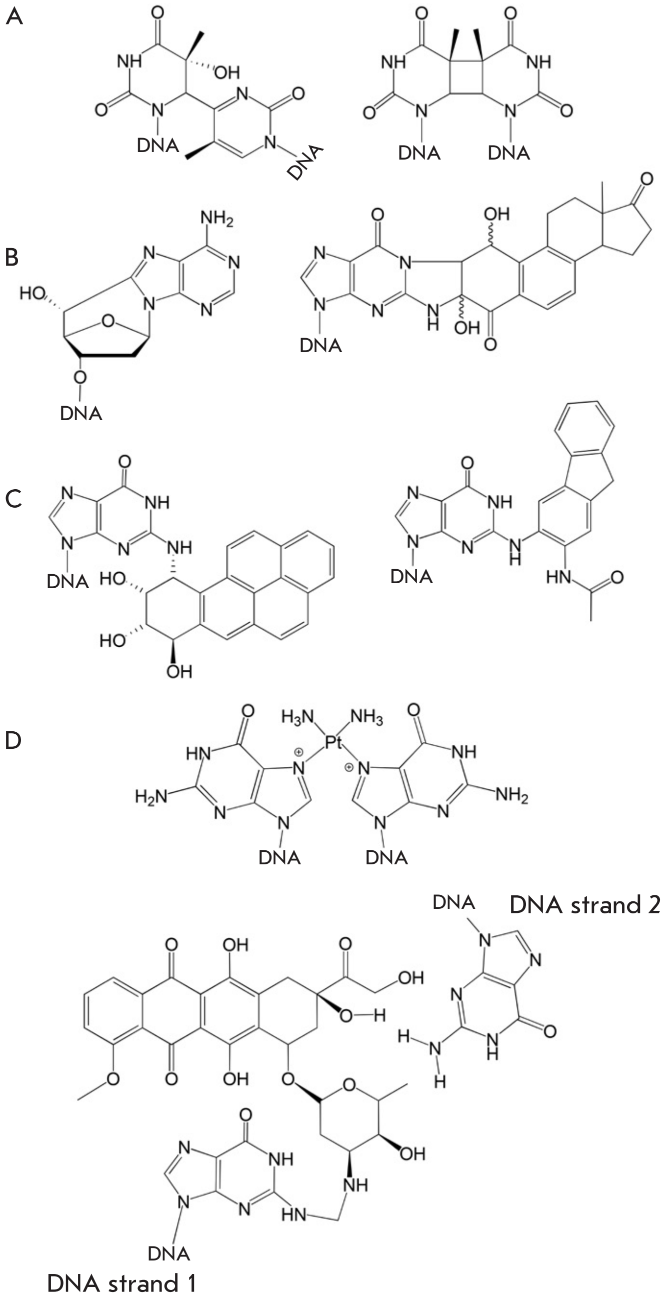


Fig. 2. Examples of DNA lesions removed by the NER system. (*A*) UV-induced lesions: a pyrimidine-(6,4)-pyrimidine photoproduct (left) and a cyclobutene pyrimidine dimer (right). (*B*) IR-induced lesions: 8,5′-cyclo-2′-deoxyadenosine (left) and a 4-hydroxyequilenin-guanine adduct (right). (*C*) DNA modifications induced by reactive environmental molecules: a benzo[a]pyrene diol epoxide-guanine adduct (left) and a (pyridyloxobutyl)guanine adduct (right). (*D*) Chemotherapy-induced lesions: a cisplatin-DNA adduct (top) and a doxorubicin-DNA adduct (bottom)

quisition of quantitative data requires the use of an isotopically labeled internal standard to allow for the formation and loss of lesions during sample processing [37–41].

In some cases, quantification results are discrepant, which may be due to both the imperfection of the used techniques and the structural features of the explored lesions [42]. These discrepancies are very typical of samples from patient tissues, tumor tissues, grafted tumors, cultured patient cells, and patient liquid biopsies, especially in cases of comprehensive (combination) therapy [25]. Further improvement of the methods for the quantification of DNA lesions is important both for identifying undesirable toxic effects on a living organism's DNA and for gaining therapeutic effect data in terms of the amount of persistent DNA lesions.

THE MECHANISMS OF INDUCTION OF CLUSTERED LESIONS CONTAINING BULKY ADDUCTS

According to rough estimates, 10^4–10^6 lesions are formed daily in the human cellular DNA [12]. Therefore, only ~0.0002–0.02% of the human genome is damaged. However, DNA lesions are nonuniformly distributed throughout the genome and are often concentrated at specific positions called mutation hotspots. Their location is indicative of both the properties of the mutation process (the predominant mutagen; efficiency of repair and replication machineries) and the structural and functional features of the cellular DNA [43].

The severity of a lesion in certain genome regions is related to many factors: the structure and amount of chemically active molecules to which the body is exposed, the mechanism of interaction between these molecules and DNA, the nucleotide sequence and local structure of DNA, and the level of chromatin compaction [43]. The small molecule–DNA interaction modes include intercalation, insertion into the minor and major DNA grooves, binding to single-stranded DNA regions, combinations of different interactions, and subsequent formation of covalent adducts with nitrogenous nucleotide bases [44].

Many substances inducing NER-repairable adducts are electrophilic compounds that interact with the nucleophilic atoms in DNA. The most reactive sites are the guanine positions N7, N2, C8, and O6; adenine positions N1, N3, and N7; thymine positions O2 and O4; and cytosine positions O2 and N4 [45]. For example, benzo[a]pyrene-7,8-diol-9,10-epoxide preferentially reacts with the guanine exocyclic (N2) amino group in the minor DNA groove. The difficult-to-repair benzo[a]pyrene adducts in this location are supposed to be the ones most often found in mammalian cellular DNA [46]. An activated aflatoxin B1 metabolite, aflatoxin B1 exo-8,9-epoxide, preferentially interacts with dG:dC-rich DNA regions and forms an adduct with (N7) guanine [47, 48]. The well-known carcinogenic aromatic amine N-2-acetylaminofluorene forms adducts at the (C8) position of guanine [49, 50]. Following metabolic activation, platinum-based chemotherapeutic agents preferentially interact with dG-rich DNA regions [51].

The risk of clustered DNA lesions significantly increases in cells under severe exposure, e.g., during intensive chemotherapy and combination therapy including exposure to radiation or additional chemotherapy drugs [5, 52, 53]. Most often, combination therapy protocols are used when essential drugs are platinum derivatives whose use is usually associated with congenital or acquired resistance. In these cases, combination therapy may include antimitotic agents terminating nucleoside analogs, topoisomerase inhibitors, and recent drugs such as paclitaxel, hemicitabine, and doxorubicin, which preferentially intercalates at the dG:dC-rich sites and forms a hydrogen bond with dG on one strand and, in the presence of formaldehyde, covalent adducts with dG on the opposite strand (*Fig. 2D*, bottom) [28].

Increased accumulation of oxidative lesions is characteristic of tumor [54, 55] and inflamed tissues [56]. Ionizing radiation induces DNA lesions both through direct ionization (30–40% of IR-induced lesions) and through exposure to free radicals generated during water radiolysis [57]. Exposure to γ- and X-ray radiation was found to lead to the formation of two or more AP sites, oxidized derivatives of nitrogenous bases, and DNA strand breaks within two or three turns of the DNA helix [58, 59]. Exposure to IR induces clustered lesions, such as AP sites and oxidized bases, about 4-fold more often than double-strand breaks [60, 61].

AP sites, one of the most numerous oxidative DNA lesions induced by exposure to various factors [62, 63], can exist as two forms in equilibrium: an open-ring aldehyde and a closed hemiacetal. The aldehyde form is highly reactive, which promotes the formation of additional lesions near AP sites. The reaction between the aldehyde form of an AP site and the exocyclic amino group of an adenine or guanine residue located in the opposite strand may result in dangerous DNA lesions – interstrand crosslinks (ICLs) [64]. A level of 20–40 ICLs per cell is lethal to repair-deficient mammalian cells [65]. These lesions block the separation of two DNA strands, which is required for transcription and replication. Therefore, ICLs act as absolute blockers of major cellular processes and are particularly detrimental to rapidly dividing cells. This has led to the widespread use of crosslinking agents as anticancer drugs. ICL repair pathways have not yet been definitively identified; NER proteins are believed to be involved in ICL repair in resting cells [65]. In addition, reactions of the aldehyde form of an AP site induce bulky adducts, such as intrachain crosslinks, mono-adducts, and DNA–protein crosslinks [64, 65].

The effect of radiomimetic agents, which are used in the chemotherapy of tumors, on DNA is similar to that of radiation. They promote the induction of multiple DNA lesions, such as single- and double-strand breaks and AP sites [66, 67]. One of these agents is bleomycin, a glycopeptide with pronounced cytotoxic and mutagenic properties, which is produced by *Streptomyces verticillus* bacteria. One part of the bleomycin molecule binds to the minor DNA groove and modifies its nitrogenous bases, while the other part is able to react with metal ions (e.g., Fe (II)) and oxygen and form reactive oxygen species that induce additional oxidative lesions in the adjacent DNA regions [66, 68].

Induction of clustered lesions is also affected by the accessibility of specific DNA regions to a damaging agent. Chromatin proteins protect DNA from the damaging effects of IR, free radicals, and genotoxic chemical compounds [69–71].

On the contrary, bulky lesions induce a significant local weakening of the Watson-Crick interactions and, thus, facilitate the accessibility of DNA to oxidative and other damaging agents and increase the likelihood of spontaneous glycosidic bond hydrolysis and AP site formation. Therefore, the presence of spontaneous or induced bulky adducts increases the risk of clustered lesions in the surrounding DNA region [62, 72]. For example, exposure of DNA containing platinum adducts to even low radiation doses was shown to increase the risk of clustered lesions 1.5 to 2.5-fold [73, 74]. Given that clustered lesions are often difficult to repair, this exposure during combination therapy may promote the accumulation of platinum adducts in the DNA of cancer cells, despite the fact that, in some cases, cancer cells are characterized by an increased activity of DNA repair systems [75, 76].

RECOGNITION OF DNA LESIONS BY GLOBAL GENOME NER

During the global genome NER process, the primary recognition of a DNA region containing a bulky lesion occurs without direct contact between the XPC sensor protein and the lesion [3, 77, 78]. As already noted, bulky lesions induce changes in the regular dsDNA structure, which are often accompanied by a destabilization of the molecule and the formation of mobile single-stranded regions with increased affinity for XPC. During the search for lesions, XPC moves along the DNA molecule in a repeated association-dissociation manner, forming many short-lived complexes with DNA, which allows XPC to bypass obstacles: proteins associated with DNA [79].

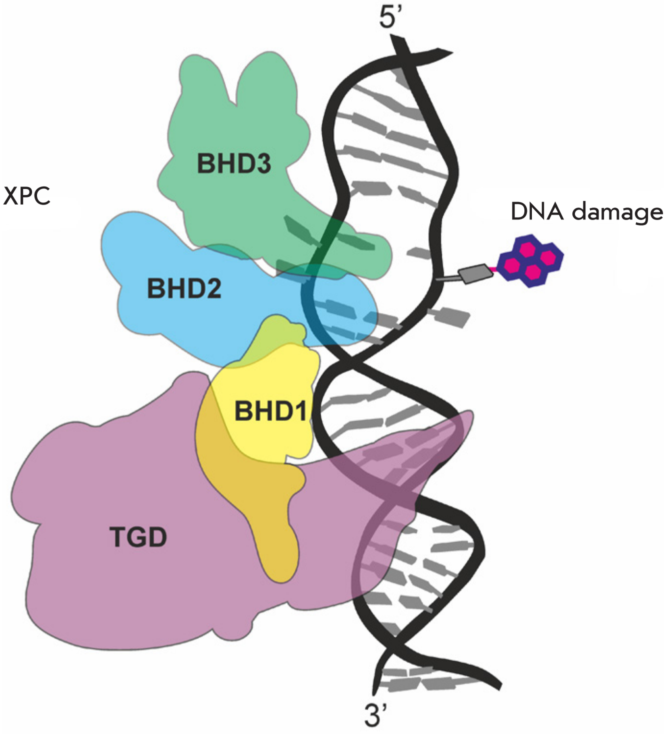


Fig. 3. DNA damage recognition by the XPC protein. DNA damage (pink), the transglutaminase (TGD) domain of XPC (purple), the BHD1 domain (yellow), the BHD2 domain (blue), and the BHD3 domain (green)

A more detailed understanding of the first NER step has been gained from biochemical experiments, such as photoaffinity modification and steady-state fluorescence titration with a recombinant human XPC protein and its yeast orthologue Rad4, as well as X-ray diffraction analysis of the Rad4 protein associated with DNA containing a cyclobutane pyrimidine dimer [77, 80]. XPC comprises three β-hairpin domains: BHD1, BHD2, and BHD3 (*Fig. 3*) [77]. At the first step of lesion recognition, the BHD1 and

BHD2 domains of the XPC factor recognize DNA regions with weakened hydrogen bonds. Regions with a weakened regular DNA duplex structure are recognized via sequential interactions between an aromatic sensor (the amino acid residues Trp690 and Phe733 located in the BHD2 domain) and aromatic heterocyclic nitrogenous bases [81, 82]. The XPC subunit comprising the aromatic sensor is similar to the oligonucleotide/oligosaccharide-binding motif typical of proteins that preferentially interact with single-stranded DNA; e.g., RPA [81–83]. The BHD1 and transglutaminase domains of XPC bind to an 11-bp segment of undamaged DNA at the 3'-end of the lesion, harboring the protein from DNA [82].

Then, a more specific XPC–DNA complex is formed in the immediate vicinity of the lesion. In this complex, two β-hairpin domains, BHD2 and BHD3 (*Fig. 3*), interact with a 4-nucleotide segment of the undamaged strand, which is located opposite the lesion (*Fig. 3*) [77, 84]. Structural studies of a complex between the yeast orthologous protein Rad4 and damaged DNA [77] revealed that binding of BHD2/3 results in the extrusion of both the damaged nucleotide and two undamaged nitrogenous bases in the complementary strand from the DNA duplex that occurs in a flipped-out open conformation. A long β-hairpin protruding from BHD3 is inserted into DNA, thereby stabilizing the structure formed during nucleotide flipping-out. In this case, the DNA backbone is kinked by about 40°. An XPC–DNA complex of a specific structure is formed, which involves a rather extended DNA region near the lesion (*Fig. 3*).

The selectivity of a search for lesions is controlled by a ratio of the time of DNA–XPC complex formation and its lifetime. Usually, NER-productive complexes are characterized by a shorter formation time and an optimal lifetime [85, 86]. Calculations performed using a model of stochastic reversible nucleoprotein NER complex formation revealed that the initial recognition of a lesion-containing DNA region is the slowest NER step that limits the rate of lesion removal [87]. The efficiency of the first NER step, recognition of damaged bases in a huge intact DNA, controls the rate of the entire repair process [85, 88, 89].

In the cell, XPC occurs as XPC–RAD23B and XPC–RAD23B–Cen2 complexes. The RAD23B subunit stabilizes the XPC protein and promotes its interaction with DNA. Following XPC binding to a damaged DNA region, the RAD23B subunit dissociates from the complex. The function of the Cen2 subunit in these complexes is not fully understood; *in vitro*, it is not required for NER [90]. However, it is known that Cen2, although not in contact with DNA, stimulates NER as a whole and is required for effective recruitment of the TFIIH factor to the repair process [91, 92].

Following the initial step of lesion recognition and XPC–DNA complex formation, a bulky DNA lesion is verified by the TFIIH factor. The TFIIH complex comprises a seven-subunit core (Core7), which is composed of the ATP-dependent helicases XPB and XPD and non-enzymatic subunits p62, p52, p44, p34, and p8, and the so-called CDK-activating kinase (CAK) complex that involves the MAT1, cyclin H, and Cdk7 subunits [93, 94]. In the presence of the CAK complex, XPB, and XPD subunits are connected via a long α-helix of the MAT1 protein, with TFIIH being in a rigid ring-like conformation that limits their enzymatic activity. After recruitment of TFIIH to NER, the CAK heterotrimer is released from the complex and Core7 forms a more flexible horseshoe-shaped structure, with XPB and XPD being located at each end of the horseshoe (*Fig. 1*) [8, 95].

Core7 binds to the repair complex through the interaction between its XPB and p62 subunits and the XPC factor associated with a damaged DNA region [96, 97]. The interaction between the XPB subunit and the XPC C-terminus located at the 5'-end from the lesion stimulates the ATPase activity of XPB and leads to the conformational rearrangement of Core7 and its binding to a DNA substrate [98, 99]. This conformational rearrangement enables XPD to bind to the damaged DNA strand on the 5'-side from the lesion.

XPD acts as a molecular sensor that verifies a bulky lesion in a DNA strand. Due to the 5'-3'-helicase activity stimulated by the p44 subunit, the protein moves to the lesion and forms an asymmetric bubble. During XPD activity, the damaged strand passes through a pore formed by the FeS, Arch, and HD1 domains of XPD and each base of the strand comes into contact with a sensor pocket on the protein surface. When XPD encounters damage, its helicase activity is inhibited and XPD is immobilized on DNA, thus marking the damage for its subsequent removal by the proteins of the incision complex (*Fig. 1*) [100, 101].

THE INFLUENCE OF THE DNA STRUCTURE ON THE REPAIR OF CLUSTERED BULKY LESIONS

Significant progress in understanding the recognition and removal of clustered lesions by the NER system has been achieved thanks to *in vitro* studies using synthetic oligodeoxyribonucleotides with lesions at specified positions of DNA strands [10, 11, 102]. *Figure 4* presents a schematic of DNA containing clustered lesions of various structures: in particular natural and synthetic bulky lesions used in these studies.

There is a direct correlation between the efficiency of the repair of some bulky DNA lesions by NER and the affinity of the XPC–RAD23B factor for these DNAs: e.g., single aminofluorene adducts located in DNA of the same sequence [88]. However, increased affinity of XPC–RAD23B for bulky lesion-containing DNAs is not always associated with a high efficiency of their excision in the case of both single, bulky lesions and clustered lesions [10, 11, 63, 86, 103]. For example, a benzo[a]pyrene adduct, *R*-cis-B[a]P-dG, is removed by NER proteins 5-fold more efficiently than the *S*-trans-B[a]P-dG isomer, despite the fact that the affinity of XPC–RAD23B for the corresponding DNA duplexes is the same [103]. Also, with minimal differences in the affinity of XPC-RAD23B for DNA with single, synthetic lesion analogs nAnt (a non-nucleotide insert with a bulky anthracenylcarbamoyl substituent) and Fap-dC (cytosine with a fluoro-chloro-azidopyridyl group introduced at the exocyclic nitrogen), the former lesion is repaired by NER proteins, while the latter is unrepairable [104].

DNAs containing clustered lesions, in particular bulky adducts, are usually characterized by increased affinity of the XPC factor for them. However, repair of such lesions by NER is partially or completely inhibited in many cases [10, 11, 63]. In [86], the real-time monitoring surface plasmon resonance technique was used to investigate interactions between XPC–RAD23B and DNAs containing single and cluster adducts formed by active metabolites of a fluorinated acetylaminofluorene derivative and C8 guanine. These adducts, forming a clustered lesion, were located in the same DNA strand and were separated by two or fewer nucleotides (*Fig. 4A*). The XPC factor was shown to form significantly more stable complexes with DNA containing clustered lesions compared with DNA containing single acetylaminofluorene adducts. In this case, NER excision activity towards DNAs with clustered lesions was lower than that towards DNAs with single lesions (in some cases, it was lower than the detection ceiling). Inhibition of specific excision in this case is supposed to be the result of disturbances in the assembly of the protein complexes responsible for the verification of DNA damage, which is due to extremely strong binding of XPC to the damaged site [86]. At K_D values of 10^{-11}–10^{-12} M, the XPC factor can compete for binding even with a single-stranded DNA sensor, the RPA protein, which, to-

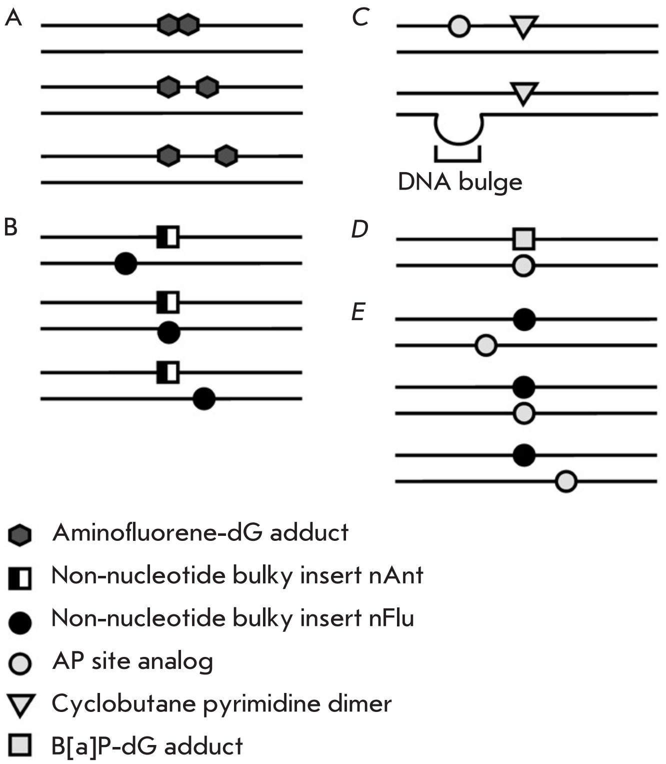

A
C
DNA bulge
B
D
E
⬢ Aminofluorene-dG adduct
◧ Non-nucleotide bulky insert nAnt
● Non-nucleotide bulky insert nFlu
○ AP site analog
▽ Cyclobutane pyrimidine dimer
□ B[a]P-dG adduct

Fig. 4. Schematic of model DNAs containing clustered lesions. (*A*) Circular plasmid DNA containing fluorinated aminofluorene mono- or di-adducts separated by one or two nucleotides. (*B*) DNA containing synthetic bulky lesions in both strands: non-nucleotide inserts containing a bulky anthracenylcarbamoyl (nAnt) or fluorescein carbamoyl (nFlu) residue; the length of a model DNA duplex is 137 bp; the interlesion distance is ≤ 20 bp. (*C*) DNA duplexes (~200 bp) containing CPD and an AP site analog in the same strand or CPD and a bulge in the complementary strand. (*D*) A 135-bp DNA duplex containing a benzo[a]pyrene diol epoxide-guanine adduct and an opposite AP site analog. (*E*) DNA containing an nFlu bulky lesion and an AP site analog in the opposite strand; the length of a model DNA duplex is 137 bp; the interlesion distance is ≤ 6 bp

gether with XPA, is part of the NER pre-incision and incision complexes [3, 105–107].

NER activity is also hindered by synthetic lesion analogs located in both DNA strands, whose bulky fragments are connected to the DNA backbone by extended flexible linkers (*Fig. 4B*) [102]. These linkers allow bulky aromatic groups of adducts to come into contact with the DNA regions adjacent to the lesion, which may induce additional destabilized DNA regions that stimulate XPC binding. A site with weakened Watson–Crick pairing near the DNA damage is supposed to be able both to inhibit and to enhance the efficiency of NER, depending on its location. The presence of this destabilization site on the 3'-side from the lesion may induce a DNA–XPC complex unproductive for subsequent NER steps: in this case, the encounter of TFIIH with the lesion is excluded [104, 108–110]. On the contrary, a DNA destabilization site on the 5'-side from the lesion may stimulate the NER process. For example, introduction of an AP site analog shifted relative to the CPD position towards the 5'-end of a damaged DNA strand was shown to stimulate excision of the CPD-containing fragment by NER [111]. A bulge in the DNA duplex on the 5'-side from CPD also increases the efficiency of its excision manifold (model DNAs are schematically shown in *Fig. 4C*). The observed effects are also associated with the features of the mechanism of lesion recognition by the TFIIH factor; namely, with the 5'-3'-direction of its movement along DNA from the primary binding site and strand unwinding direction.

Of particular interest is the investigation of the repair mechanism of clustered lesions composed of bulky DNA adducts and oxidative lesions to DNA nitrogenous bases [10, 11]. As mentioned above, a DNA region destabilized by a bulky lesion is more susceptible to reactive oxygen species, thus increasing the risk of clustered lesions. These clustered lesions can attract nucleotide excision repair and base excision repair (BER) proteins.

The repair of a clustered lesion composed of a bulky B[a]P adduct and an AP site analog, which are located in the complementary strands of the DNA duplex, was analyzed in [10] (model DNA is shown in *Fig. 4D*). Evaluation of the NER excision activity towards B[a]P-dG and the ability of AP endonuclease 1 to hydrolyze the AP site showed that NER was inhibited in these clusters, while the AP sites were repaired by BER. Therefore, the NER system is sensitive to oxidative AP site-like lesions in the immediate vicinity of B[a]P-dG [8, 10]. A further detailed study of the interaction between this model structure and repair proteins revealed that XPC stimulated the endonuclease activity and inhibited the 3'-5'-exonuclease activity of AP-endonuclease 1, thereby increasing the efficiency of BER [63].

Liu et al. [10] used NMR spectroscopy, measurements of the DNA duplex thermal stability, and computer simulation to demonstrate that DNA containing an AP site opposite a B[a]P-dG adduct is characterized by strong stacking interactions between B[a]P aromatic rings and neighboring nitrogenous bases of the complementary strand, which may inhibit XPC–DNA complex formation. In this case, the flipping of neighboring nucleotides, insertion of a β-hairpin of the BHD3 domain, and extrusion of the lesion from the DNA helix are impeded. Moreover, the XPC factor was characterized by increased affinity for the tested DNAs a containing clustered lesion [63].

A benzo[a]pyrene adduct also became unrepairable by NER upon deletion of its complementary dC nucleotide. NMR spectroscopy and computer simulation studies [9, 112] demonstrated that deletion of dC significantly enhances stacking interactions between the B[a]P aromatic ring and the surrounding nitrogenous bases, which prevents the formation of a productive open XPC–DNA complex [112].

Naumenko et al. [11] explored the effect of an AP site analog located on different sites of the complementary DNA strand on the removal of a non-nucleotide insert comprising a bulky fluorescein carbamoyl fragment (nFlu) by NER (*Fig. 4E*). The XPC factor and DNA formed unproductive complexes in which the nFlu bulky lesion and the AP site analog were separated by less than 6 bps. There was an inverse correlation between the relative efficiency of excision of nFlu-containing fragments from these model DNAs and the affinity of XPC for the model DNAs. The location of the AP site and nFlu in opposite positions of the DNA duplex, as well as similar localization of other lesions (B[a]P-dG/AP site, nAnt/nFlu), completely inhibited the excision of the bulky damage by NER proteins (*Fig. 4B, D*) [10, 102].

Structural DNA changes associated with inhibited nFlu excision in the presence of an AP site analog in the complementary strand (*Fig. 4E*) were revealed using molecular dynamics. Simulation of molecular dynamics trajectories showed that DNA with nFlu and an AP site analog, which were located opposite each other in the complementary strands, was in a “compressed” conformation of the duplex at the lesion site: the bases adjacent to the lesion were characterized by effective stacking interactions with each other most of the time, and both lesions were flipped out of the strands. The fluorescein moiety (Flu) occurred in the minor groove, oriented towards the 5'-end of the damaged strand, which had the potential to sterically hinder binding of XPC to a destabilized DNA region located on the 5'-side from the lesion. In this case, an unproductive XPC binding site on the 3'-side from nFlu became more accessible [11], which has the potential to lead to the formation of an XPC–DNA complex unproductive for NER. For a short time, Flu may be oriented towards the 3'-end of the damaged strand and interact with the AP site analog on the opposite side of the DNA helix.

Therefore, an additional non-bulky lesion of a nitrogenous base (e.g., an AP site) or a deletion in the complementary strand, opposite a bulky DNA adduct, may induce local stabilization of the damaged site [9–11, 112], which prevents binding of the XPC factor, thus excluding the subsequent NER steps.

The verification step may also affect the efficiency of NER lesion removal. Affinity of XPD for model DNAs containing single bulky lesions with similar XPC affinity (K_D = 1.5–3 nM) was recently shown to depend on the structure of bulky lesions and vary significantly, being correlated with the efficiency of lesion removal *in vitro* [104, 113]. The number of studies on the verification of clustered DNA lesions is rather small. For example, introduction of an AP site into a DNA substrate (either into the strand scanned by XPD helicase or into the complementary strand, "invisible" for XPD) was shown not to significantly affect the helicase and ATPase activity of recombinant Core7 [106]. Thus, the verification step is unlikely to promote significant differences in the efficiency of NER in DNAs containing clustered lesions of this composition.

It should be noted that obstacles to a successful repair of a bulky adduct from a clustered lesion may also include steric hindrances during excision of a damaged DNA fragment by the XPF and XPG endonucleases and the lack of an undamaged DNA template of the complementary strand. However, this topic has not been well addressed and requires further research.

CONCLUSION

Due to differences in the chemical properties of nitrogenous DNA bases and the type and strength of genotoxic factors, lesions are unevenly distributed over cellular DNA, concentrating in certain regions of the genome. Clustered lesions are often difficult to repair, which leads to their accumulation in DNA, especially if the repair status of the cell is reduced. On the other hand, hindered DNA repair of induced lesions should promote their cytotoxic effect on cancer cells. Model DNA studies have shown that removal of bulky lesions during global genome NER may be inhibited at the initial recognition step due to the structural features of a cluster-containing DNA region. In the clustered DNA lesions formed by a bulky adduct and an opposite AP site, the AP site was shown to be processed by BER enzymes rather efficiently, while NER excision of a bulky lesion from these structures was difficult. Sequential removal of lesions from clusters is supposed to be of adaptive value, because it excludes simultaneous initiation of NER and BER. Understanding the mechanisms of removal of clustered DNA lesions containing bulky adducts should help develop rational and efficient approaches to the maintenance of therapy-induced DNA lesions in cancer cells with increased activity of DNA repair systems. ●

This work was supported by the Russian Science Foundation (grant No. 19-74-10056) and state assignment for the Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences (No. 121031300041-4, section "ORIGIN AND TYPES OF NER-REPAIRABLE DNA LESIONS").

REFERENCES

1. Fousteri M., Mullenders L.H.F. // Cell Res. 2008. V. 18. № 1. P. 73–84.
2. Vermeulen W., Fousteri M. // Cold Spring Harb. Perspect. Biol. 2013. V. 5. № 8. P. 1–16.
3. Schärer O.D. // Cold Spring Harb. Perspectives Biol. 2013. V. 5. № 10. P. 1–19.
4. Petruseva I.O., Evdokimov A.N., Lavrik O.I. // Acta Naturae. 2014. V. 6. № 20. P. 23–34.
5. Sage E., Harrison L. // Mutat. Res. – Fundam. Mol. Mech. Mutagen. 2011. V. 711. № 1–2. P. 123–133.
6. Eccles L.J., O'Neill P., Lomax M.E. // Mutat. Res. – Fundam. Mol. Mech. Mutagen. 2011. V. 711. № 1–2. P. 134–141.
7. Sage E., Shikazono N. // Free Radic. Biol. Med. 2017. V. 107. P. 125–135.
8. Mu H., Geacintov N.E., Broyde S., Yeo J.E., Schärer O.D. // DNA Repair (Amst.). 2018. V. 71. P. 33–42.
9. Hess M.T., Gunz D., Luneva N., Geacintov N.E., Naegeli H. // Mol. Cell. Biol. 1997. V. 17. № 12. P. 7069–7076.
10. Liu Z., Ding S., Kropachev K., Lei J., Amin S., Broyde S. // PLoS One. 2015. V. 10. № 9. e0137124.
11. Naumenko N., Petruseva I., Lomzov A., Lavrik O. // DNA Repair (Amst.). 2021. V. 108. 103225.
12. Schärer O.D. // Angew. Chem. Int. Ed. 2003. V. 42. P. 2946–2974.
13. Lloyd D.R., Hanawalt P.C. // Cancer Res. 2002. V. 62. № 18. P. 5288–5294.
14. Duan M., Ulibarri J., Liu K.J., Mao P. // Int. J. Mol. Sci. 2020. V. 21. № 23. P. 1–13.
15. Rastogi R.P., Kumar A., Tyagi M.B., Sinha R.P. // J. Nucl. Acids. 2010. 592980.
16. Brooks P.J., Wise D.S., Berry D.A., Kosmoski J.V., Smerdon M.J., Somers R.L., Mackie H., Spoonde A.Y., Ackerman E.J., Coleman K., et al. // J. Biol. Chem. 2000. V. 275. № 29. P. 22355–22362.
17. Okahashi Y., Iwamoto T., Suzuki N., Shibutani S., Sugiura S., Itoh S., Nishiwaki T., Ueno S., Mori T. // Nucl. Acids Res. 2010. V. 38. № 12. e133.
18. Pratt M.M., John K., Maclean A.B., Afework S., Phillips D.H., Poirier M.C. // Int. J. Environ. Res. Public Hlth. 2011. V. 8. № 7. P. 2675–2691.
19. Phillips D.H. // Environ. Health Perspect. 1996. V. 104. № 3. P. 453–458.
20. Guo S., Leng J., Tan Y., Price N.E., Wang Y. // Chem. Res. Toxicol. 2019. V. 32. № 4. P. 708–717.
21. Baker D.J., Wuenschell G., Xia L., Termini J., Bates S.E., Riggs A.D., O'Connor T.R. // J. Biol. Chem. 2007. V. 282. № 31. P. 22592–22604.
22. Grollman A.P., Shibutani S., Moriya M., Miller F., Wu L., Moll U., Suzuki N., Fernandes A., Rosenquist T., Medverec Z., et al. // Proc. Natl. Acad. Sci. USA. 2007. V. 104. № 29. P. 12129–12134.
23. Sidorenko V.S., Yeo J.E., Bonala R.R., Johnson F., Schärer O.D., Grollman A.P. // Nucl. Acids Res. 2012. V. 40. № 6. P. 2494–2505.
24. Geacintov N.E., Broyde S. // Chem. Res. Toxicol. 2017. V. 30. P. 1517–1548.
25. Stornetta A., Zimmermann M., Cimino G.D., Henderson P.T., Sturla S.J. // Chem. Res. Toxicol. 2017. V. 30. № 1. P. 388–409.
26. Gong J., Vaidyanathan V.G., Yu X., Kensler T.W., Peterson L.A., Sturla S.J. // J. Am. Chem. Soc. 2007. V. 129. № 7. P. 2101–2111.
27. Boamah E.K., Brekman A., Tomasz M., Myeku N., Figueiredo-Pereira M., Hunter S., Meyer J., Bhosle R.C., Bargonetti J. // Chem. Res. Toxicol. 2010. V. 23. № 7. P. 1151–1162.
28. Yang F., Teves S.S., Kemp C.J., Henikoff S., Lecture N., Gasser H., Yang F., Teves S.S., Kemp C.J., Henikoff S. // Biochim. Biophys. Acta – Rev. Cancer. 2014. V. 1845. № 1. P. 84–89.
29. Hartmann A., Agurell E., Beevers C., Brendler-Schwaab S., Burlinson B., Clay P., Collins A., Smith A., Speit G., Thybaud V., et al. // Mutagenesis. 2003. V. 18. № 1. P. 45–51.
30. Buss P., Caviezel M., Lutz W.K. // Carcinogenesis. 1990. V. 11. № 12. P. 2133–2135.
31. Phillips D.H., Farmer B.P., Beland F.A., Nath R.G., Poirier M.C., Reddy M.V., Turteltaub K.W. // Env. Mol. Mutagen. 2000. V. 35. № 3. P. 222–233.
32. Singh R., Gaskell M., Le Pla R.C., Kaur B., Azim-Araghi A., Roach J., Koukouves G., Souliotis V.L., Kyrtopoulos S.A., Farmer P.B. // Chem. Res. Toxicol. 2006. V. 19. № 6. P. 868–878.
33. Poirier M.C. // Environ. Health Perspect. 1997. V. 105. № 4. P. 907–912.
34. Lippard S.J., Merkel C.M., Lippard S.J., Ushay H.M., Poirier M.C., Poirier M.C. // Biochemistry. 1983. V. 22. № 22. P. 5165–5168.
35. Aloisi C.M.N., Nilforoushan A., Ziegler N., Sturla S.J. // J. Am. Chem. Soc. 2020. V. 142. № 15. P. 6962–6969.
36. Yang Z., Faustino P.J., Andrews P.A., Monastra R., Rasmussen A.A., Ellison C.D., Cullen K.J. // Cancer Chemother. Pharmacol. 2000. V. 46. № 4. P. 255–262.
37. Otteneder M., Lutz W.K. // Mutat. Res. – Fundam. Mol. Mech. Mutagen. 1999. V. 424. № 1–2. P. 237–247.
38. Farmer P.B., Brown K., Tompkins E., Emms V.L., Jones D.J.L., Singh R., Phillips D.H. // Toxicol. Appl. Pharmacol. 2005. V. 207. № 2. P. 293–301.
39. Farmer P.B., Singh R. // Mutat. Res. 2008. V. 659. № 1–2. P. 68–76.
40. Brown K. // Methods Mol. Biol. 2012. V. 817. P. 207–230.
41. Tretyakova N., Goggin M., Sangaraju D., Janis G. // Chem. Res. Toxicol. 2012. V. 25. № 10. P. 2007–2035.
42. Collins A., Gedik C., Vaughan N., Wood S., White A., Dubois J., Rees J.F., Loft S., Møller P., Poulsen H., et al. // Free Radic. Biol. Med. 2003. V. 34. № 8. P. 1089–1099.
43. Rogozin I.B., Pavlov Y.I. // Mutat. Res. – Rev. Mutat. Res. 2003. V. 544. № 1. P. 65–85.
44. Sturla S.J. // Curr. Opin. Chem. Biol. 2007. V. 11. № 3. P. 293–299.
45. Gillet L.C., Schärer O.D. // Chem. Rev. 2006. V. 106. № 2. P. 253–276.
46. Hargis J., Schaefer H., Houk K., Wheeler S. // J. Phys. Chem. A. 2010. V. 8. № 24. P. 4017–4018.
47. Eaton D.L., Gallagher E.P. // Annu. Rev. Pharmacol. Toxicol. 1994. V. 34. № 1. P. 135–172.
48. Bedard L.L., Massey T.E. // Cancer Lett. 2006. V. 241. № 2. P. 174–183.
49. Romano L., Vooradi V. // Biochemistry. 2010. V. 38. № 3. P. 319–335.
50. Mu H., Kropachev K., Wang L., Zhang L., Kolbanovskiy A., Kolbanovskiy M., Geacintov N.E., Broyde S. // Nucl. Acids Res. 2012. V. 40. № 19. P. 9675–9690.
51. Hemminki K., Thilly W.G. // Mutat. Res. – Fundam. Mol. Mech. Mutagen. 1988. V. 202. № 1. P. 133–138.
52. Mohamad O., Sishc B.J., Saha J., Pompos A., Rahimi A., Story M.D., Davis A.J., Kim D.W.N. // Cancers (Basel). 2017. V. 9. № 6. P. 1–30.
53. Regulus P., Duroux B., Bayle P.A., Favier A., Cadet J.,

Ravanat J.L. // Proc. Natl. Acad. Sci. USA. 2007. V. 104. № 35. P. 14032–14037.
54. Iida T. // Neuro. Oncol. 2001. V. 3. № 2. P. 73–81.
55. Goodman M., Bostick R.M., Dash C., Terry P., Flanders W.D., Mandel J. // Cancer Causes Control. 2008. V. 19. № 10. P. 1051–1064.
56. Grivennikov S.I., Greten F.R., Karin M. // Cell. 2010. V. 140. № 6. P. 883–899.
57. Ward J.F. // Radiat. Res. 1994. V. 138. № 1 (suppl). P. S85-S88.
58. Goodhead D.T. // Int. J. Radiat. Biol. 1994. V. 65. № 1. P. 7–17.
59. Watanabe R., Rahmanian S., Nikjoo H. // Radiat. Res. 2015. V. 183. № 5. P. 525–540.
60. Sutherland B.M., Bennett P.V., Sidorkina O., Laval J. // Mutagenesis. 2000. V. 97. № 1. P. 247–261.
61. Gulston M., de Lara C., Jenner T., Davis E., O'Neill P. // Nucl. Acids Res. 2004. V. 32. № 4. P. 1602–1609.
62. Gates K.S. // Chem. Res. Toxicol. 2009. V. 22. № 11. P. 1747–1760.
63. Starostenko L.V., Maltseva E.A., Lebedeva N.A., Pestryakov P.E., Lavrik O.I., Rechkunova N.I. // Biochem. (Moscow). 2016. V. 81. № 3. P. 233–241.
64. Greenberg M.M. // Acc. Chem. Res. 2014. V. 47. № 2. P. 646–655.
65. Clauson C., Schärer O.D., Niedernhofer L. // Cold Spring Harb. Perspect. Biol. 2013. V. 5. № 10. a012732.
66. Chen J., Stubbe J.A. // Nat. Rev. Cancer. 2005. V. 5. № 2. P. 102–112.
67. Nickoloff J.A., Sharma N., Taylor L. // Genes. 2020. V. 11. № 1. P. 99.
68. Smith B.L., Bauer G.B., Povirk L.F. // J. Biol. Chem. 1994. V. 269. № 48. P. 30587–30594.
69. Takata H., Hanafusa T., Mori T., Shimura M., Iida Y., Ishikawa K., Yoshikawa K., Yoshikawa Y., Maeshima K. // PLoS One. 2013. V. 8. № 10. e75622.
70. Ljungman M., Hanawalt P.C. // Mol. Carcinog. 1992. V. 5. № 4. P. 264–269.
71. Falk M., Lukášová E., Kozubek S. // Biochim. Biophys. Acta – Mol. Cell Res. 2008. V. 1783. № 12. P. 2398–2414.
72. Gunz D., Hess M.T., Naegeli H., For E., Probing A.T. // J. Biol. Chem. 1996. V. 271. № 41. P. 25089–25098.
73. Dong Y., Chen Y., Zhou L., Shao Y., Fu X., Zheng Y. // Int. J. Radiat. Biol. 2017. V. 93. № 11. P. 1274–1282.
74. Zheng Y., Sanche L. // Int. J. Mol. Sci. 2019. V. 20. № 15. 3749.
75. Cheung-Ong K., Giaever G., Nislow C. // Chem. Biol. 2013. V. 20. № 5. P. 648–659.
76. Fu D., Calvo J.A., Samson L.D. // Nat. Rev. Cancer. 2012. V. 12. № 2. P. 104–120.
77. Min J., Pavletich N.P. // Nature. 2007. V. 449. P. 570–575.
78. Jain V., Hilton B., Lin B., Patnaik S., Liang F., Darian E., Zou Y., MacKerell A.D., Cho B.P. // Nucl. Acids Res. 2013. V. 41. № 2. P. 869–880.
79. Cheon N.Y., Kim H.S., Yeo J.E., Schärer O.D., Lee J.Y. // Nucl. Acids Res. 2019. V. 47. № 16. P. 8337–8347.
80. Krasikova Y.S., Rechkunova N.I., Maltseva E.A., Pestryakov P.E., Petruseva I.O., Sugasawa K., Chen X., Min J.H., Lavrik O.I. // J. Biol. Chem. 2013. V. 288. № 15. P. 10936–10947.
81. Maillard O., Solyom S., Naegeli H. // PLoS Biol. 2007. V. 5. № 4. e79.
82. Camenisch U., Träutlein D., Clement F.C., Fei J., Leitenstorfer A., Ferrando-May E., Naegeli H. // EMBO J. 2009. V. 28. № 16. P. 2387–2399.
83. Bunick C.G., Miller M.R., Fuller B.E., Fanning E., Chazin W.J. // Biochemistry. 2006. V. 45. P. 14965–14979.
84. Paul D., Mu H., Zhao H., Ouerfelli O., Jeffrey P.D., Broyde S., Min J.H. // Nucl. Acids Res. 2019. V. 47. № 12. P. 6015–6028.
85. Chen X., Velmurugu Y., Zheng G., Park B., Shim Y., Kim Y., Liu L., van Houten B., He C., Ansari A., et al. // Nat. Commun. 2015. V. 6. 5849.
86. Hilton B., Gopal S., Xu L., Mazumder S., Musich P.R., Cho B.P., Zou Y.Z. // PLoS One. 2016. V. 11. № 6. P. 1–21.
87. Luijsterburg M.S., Bornstaedt G. Von, Gourdin A.M., Politi A.Z., Moné M.J., Warmerdam D.O., Goedhart J., Vermeulen W., Driel R. Van, Höfer T. // J. Cell Biol. 2010. V. 189. № 3. P. 445–463.
88. Yeo J.E., Khoo A., Fagbemi A.F., Schärer O.D. // Chem. Res. Toxicol. 2012. V. 25. № 11. P. 2462–2468.
89. Buterin T., Meyer C., Giese B., Naegeli H. // Chem. Biol. 2005. V. 12. № 8. P. 913–922.
90. Nishi R., Sakai W., Tone D., Hanaoka F., Sugasawa K. // Mol. Cell. Biol. 2013. V. 41. № 14. P. 5664–5674.
91. Dantas T.J., Wang Y., Lalor P., Dockery P., Morrison C.G. // J. Cell Biol. 2011. V. 193. № 2. P. 307–318.
92. Bergink S., Toussaint W., Luijsterburg M.S., Dinant C., Alekseev S., Hoeijmakers J.H.J., Dantuma N.P., Houtsmuller A.B., Vermeulen W. // J. Cell Biol. 2012. V. 196. № 6. P. 681–688.
93. Volker M., Moné M.J., Karmakar P., Van Hoffen A., Schul W., Vermeulen W., Hoeijmakers J.H.J., van Driel R., van Zeeland A.A., Mullenders L.H.F. // Mol. Cell. 2001. V. 8. № 1. P. 213–224.
94. Araújo S.J., Nigg E.A., Wood R.D. // Mol. Cell. Biol. 2001. V. 21. № 7. P. 2281–2291.
95.Greber B.J., Hoang T., Nguyen D., Fang J., Afonine P.V., Paul D., Nogales E., Division I.B., National L.B., Biology C. // Nature. 2017. V. 549. № 7672. P. 414–417.
96. Yokoi M., Masutani C., Maekawa T., Sugasawa K., Ohkuma Y., Hanaoka F. // J. Biol. Chem. 2000. V. 275. № 13. P. 9870–9875.
97. Greber B.J., Toso D.B., Fang J., Nogales E. // Elife. 2019. V. 8. e44771.
98. Houten B., van Kuper J., Kisker C. // DNA Repair (Amst.). 2016. V. 44. P. 136–142.
99. Oksenych V., Coin F. // Cell Cycle. 2010. V. 9. № 1. P. 90–96.
100. Egly J.M., Coin F. // DNA Repair (Amst.). 2011. V. 10. № 7. P. 714–721.
101. Kuper J., Braun C., Elias A., Michels G., Sauer F., Schmitt D.R., Poterszman A., Egly J.M., Kisker C. // PLoS Biol. 2014. V. 12. № 9. e1001954.
102. Lukyanchikova N.V., Petruseva I.O., Evdokimov A.N., Koroleva L.S., Lavrik O.I. // Mol. Biol. 2018. V. 52. № 2. P. 237–246.
103. Lee Y.C., Cai Y., Mu H., Broyde S., Amin S., Chen X., Min J.H., Geacintov N.E. // DNA Repair (Amst.). 2014. V. 19. P. 55–63.
104. Evdokimov A.N., Tsidulko A.Y., Popov A.V., Vorobiev Y.N., Lomzov A.A., Koroleva L.S., Silnikov V.N., Petruseva I.O., Lavrik O.I. // DNA Repair (Amst.). 2018. V. 61. P. 86–98.
105. Krasikova Y.S., Rechkunova N.I., Maltseva N.I., Petruseva I.O., Lavrik O.I. // Nucl. Acids Res. 2010. V. 38. № 22. P. 8083–8094.
106. Li C.L., Golebiowski F.M., Onishi Y., Samara N.L., Sugasawa K., Yang W. // Mol. Cell. 2015. V. 59. № 6. P. 1025–1034.

107. Marteijn J.A., Hoeijmakers J.H.J., Vermeulen W. // Mol. Cell. 2015. V. 59. № 6. P. 885–886.
108. Sugasawa K., Shimizu Y., Iwai S., Hanaoka F. // DNA Repair (Amst.). 2002. V. 1. № 1. P. 95–107.
109. Sugasawa K., Akagi J., Nishi R., Iwai S., Hanaoka F. // Mol. Cell. 2009. V. 36. № 4. P. 642–653.
110. Nemzow L., Lubin A., Zhang L., Gong F. // DNA Repair (Amst.). 2015. V. 36. P. 19–27.
111. Kusakabe M., Onishi Y., Tada H., Kurihara F., Kusao K., Furukawa M., Iwai S., Yokoi M., Sakai W., Sugasawa K. // Genes Environ. 2019. V. 41. P. 1–6.
112. Reeves D.A., Mu H., Kropachev K., Cai Y., Ding S., Kolbanovskiy A., Kolbanovskiy M., Chen Y., Krzeminski J., Amin S., et al. // Nucl. Acids Res. 2011. V. 39. № 20. P. 8752–8764.
113. Petruseva I., Naumenko N., Kuper J., Anarbaev R., Kappenberger J., Kisker C., Lavrik O. // Front. Biosci. 2021. V. 9. 617160.

About the Biodiversity of the Air Microbiome

N. B. Naumova, M. R. Kabilov*
Institute of Chemical Biology and Fundamental Medicine, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, 630090 Russia
*E-mail: kabilov@niboch.nsc.ru

Received December 18, 2021; in final form, October 27, 2022
DOI: 10.32607/actanaturae.11671

ABSTRACT This brief review focuses on the properties of bioaerosols, presenting some recent results of metagenomic studies of the air microbiome performed using next-generation sequencing. The taxonomic composition and structure of the bioaerosol microbiome may display diurnal and seasonal dynamics and be dependent on meteorological events such as dust storms, showers, fogs, etc., as well as air pollution. The Proteobacteria and Ascomycota members are common dominants in bioaerosols in different troposphere layers. The microbiological composition of the lower troposphere air affects the composition and diversity of the indoor bioaerosol microbiome, and information about the latter is very important, especially during exacerbated epidemiological situations. Few studies focusing on the bioaerosol microbiome of the air above Russia urge intensification of such research.

KEYWORDS bioaerosol, microbiome, troposphere, atmospheric transport, biodiversity.

INTRODUCTION

Microorganisms are found ubiquitously in the environment and play a crucial role in almost all ecosystems [1]. Since many pathogens spread through the airborne route, including the SARS-CoV-2 coronavirus that has caused the current COVID-19 pandemic, it is especially relevant to study, monitor, and control the composition of outdoor and indoor air [2, 3]. Much data have been gained about the correlation between outdoor air pollution and the more severe course of COVID-19: for example, in India, a lower mortality rate from COVID-19 was observed in cities with better air quality [4]. We would like to emphasize that the term “bioaerosol” covers a broad range of particulate organic matter contained in the atmosphere, originating from various living and dead organisms [5]. Along with particulate matter of microbial, plant, or animal origin, bioaerosols usually also contain a broad range of antigenic compounds, microbial toxins, and viruses [6, 7]. Understanding the processes of bioaerosol formation, their distribution patterns, migration, structure, etc., especially under the harsh conditions of the upper atmosphere, is required for many fundamental and applied scientific disciplines [8], such as physics, chemistry, meteorology, and atmospheric hydrology; research into the content of allergy-inducing particles and microorganisms pathogenic to humans, farm animals, and plants; as well as aerobiology, biogeography, biodiversity, and general ecology. The key trends in bioaerosol research include (a) assessment of their sources and flows, (b) spatial distribution and its changes over time, (c) aging of biological particles, (d) metabolic activity, (e) urbanization of allergies, (f) pathogen transport, and (g) the impact on climate [8].

This review aims to briefly describe the bioaerosol microbiota, with special focus placed on the microbiome composition and structure. Air is an extremely dynamic (and, therefore, very challenging) environment for collecting and analyzing bioaerosol samples, identifying the aerosolization sources and transport pathways, so the methodological aspects of sample collection are undoubtedly of great significance for data interpretation and comparison. The microbiome analysis techniques are also very important. Nevertheless, since these two trends are rather extensive, we will touch upon them only briefly in this review.

THE MAIN PROPERTIES OF BIOAEROSOLS

Bioaerosols are an important component of atmospheric aerosols. Calculations show that bioaerosols account for 10–28 vol.% [9] and 16–80 wt.% of all the particulate matter found in the air [1].

The airborne transmission of microorganisms is ubiquitous, being an essential part of the life cycle for some of them [10]. Various natural sources such as

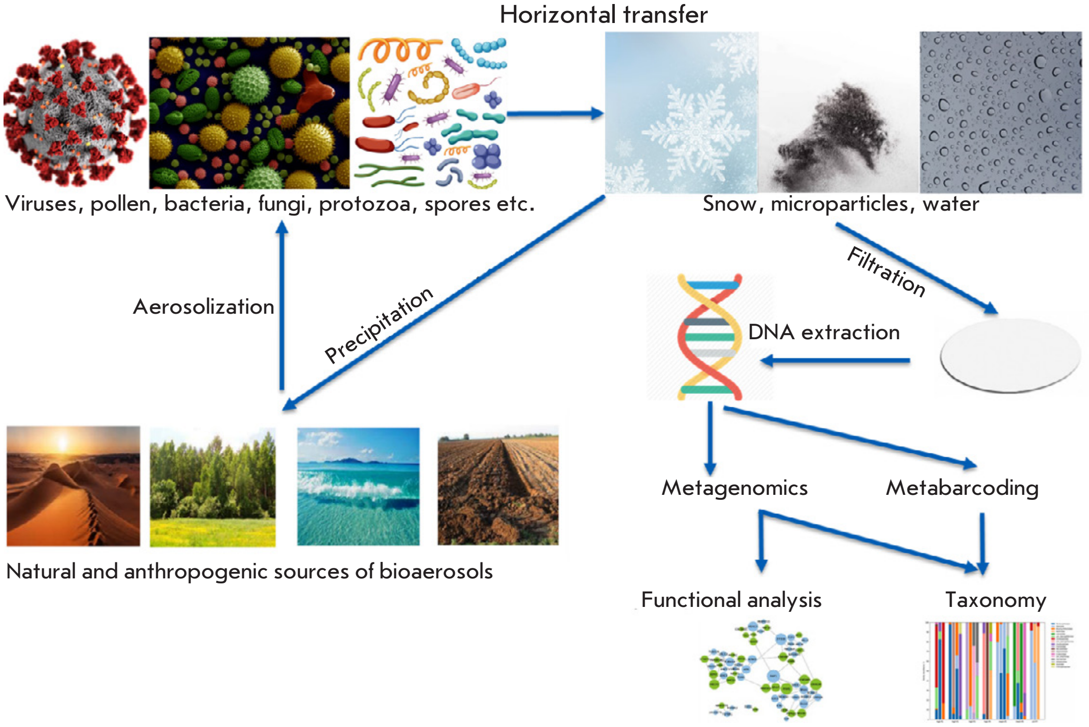

Horizontal transfer
Viruses, pollen, bacteria, fungi, protozoa, spores etc.
Snow, microparticles, water
Aerosolization
Precipitation
Filtration
DNA extraction
Natural and anthropogenic sources of bioaerosols
Metagenomics
Metabarcoding
Functional analysis
Taxonomy

Fig. 1. Bioaerosol formation, distribution, and analysis: schematic presentation

soil, forests, deserts, oceans, seas, etc. [11], as well as anthropogenic ones (agriculture, food industry, landfills, etc.), contribute to bioaerosol formation (*Fig. 1*) [6, 11, 12].

Once microorganisms get into the atmosphere (i.e., are aerosolized), they are much more likely to be exposed to the stress caused by drying, UV radiation, low temperatures, low carbon content, and low energy compared to their natural habitats (the sources of aerosol): so, many microorganisms may die [13].

The size of bioaerosol particles varies from 3 nm [14] to 100 μm, depending on their source: the diameter of pollen is 17–58 μm; that of fungal spores, 1–30 μm; the diameter of bacterial cells usually is 0.25–8 μm [15]; and that of viruses, < 0.3 μm. Meanwhile, the biological material does not necessarily consist of individual particles: most bacteria are associated with particles with a diameter > 2 μm [16, 17], 2–3 μm [18], and 3–4 μm [19, 20]. In some cases, bacteria were found to be characterized by a bimodal bioaerosol particle size distribution with the peaks at 1–2 μm and 4–7 μm [21]. Bacteria can also occur as cell agglomerates or be associated with plant, animal, or soil particles, as well as pollen or spores. Airborne bacterial cells and fungal spores can have concentrations as high as ~ $10^3 \div 10^4$ and ~10^5 per m^3 [17, 21] and be found at altitudes up to 40 km above sea level; i.e., up in the stratosphere [22]. In the near-surface layer of the troposphere, the concentration of bacterial particles capable of forming colonies on laboratory culture media ranged from 65 to 355 CFU/m^3 in urban areas in southern Poland [19] and from 300 to 1350 CFU/m^3 in urban and rural areas in Thailand [18]. In the latter case, the number of CFUs decreased rapidly with altitude (twofold when proceeding from 1–3 m to 7 m above ground level). Laboratory cultivation revealed that spore-forming bacteria Bacilli/Firmicutes were significantly dominant in the near-surface and higher layers (several thousand meters) of the troposphere over the south of West Siberia [23, 24], while non-spore-forming bacteria were dominant over the northern part of this region [25].

Bioaerosol distribution in the air depends on the particular season [19, 26]. Thus, the concentration of bacterial cells in the air of the coastal region of China determined microscopically was higher in winter than in summer [21]. The bioaerosol load with pathogenic microbiota can vary greatly depending on the time of year: in South Asia, the pathogen content was found to substantially rise during the post-monsoon season and winter. Significant diurnal variation in bioaerosol composition was also detected [26].

Temperature and ultraviolet radiation are the most statistically significant meteorological factors responsible for the viability of airborne bacteria [19, 21]. The aerosol load with biota and their behavior in the environment largely depend on air pollution (haze, fog, dust, and various macroparticles), including pollution from transportation and biomass burning [26]. The proportion of viable bacteria in the total pool depends on the degree of pollution [15]. The bioaerosol composition can vary depending on specific, random meteorological conditions: for example, dust storms strongly increased the concentration of microorganisms in bioaerosols [16], and different bioaerosol components vary in different ways depending on meteorological conditions.

The data obtained so far indicate that bioaerosols play an important role [6, 11, 27–28] in the physical and chemical processes occurring in the atmosphere [1, 29]. It was shown that bioaerosols can bind to surrounding particles, thus influencing atmospheric processes by acting as condensation nuclei in clouds and initiating precipitation [10, 30–31]. Thus, it was found that biological particles act as nuclei for snow and cloud formation in 33% of cases [32].

Along with having an impact on weather phenomena, bioaerosols also affect human health [33], since they may contain pathogenic or opportunistic bacteria, fungi, viruses, high-molecular-weight allergens, bacterial endotoxins, mycotoxins, peptidoglycans, β(1-3)-glycans, pollen, and plant fibers [6]. First, the unfavorable effects of bioaerosols on human health manifest themselves as respiratory symptoms. Thus, there is a strong correlation between the increased outdoor air pollen concentration in spring and summer and asthma exacerbation in children [34]. An association was found between the content of fungal spores in the air and the number of patients with asthma symptoms requesting medical assistance [35]. The endotoxin of bacterial bioaerosols is considered an important etiological factor of occupational lung diseases, including non-allergic asthma [6]. *Escherichia coli* isolates, which are commonly used as water quality indicators, have also been found in atmospheric dust [36].

AEROSOL SAMPLE COLLECTION

Aerosol sample collection is based on various physical approaches to separating particles from the air flow [37]. But the general idea is to pump air through a filter or a fluid medium, entrapping aerosol particles [38]. Techniques allowing one to separate particles according to their size during sample collection have recently started to appear [39]. They are particularly relevant for aerovirology: over the past decade, there has been intense research into the methods that can be used to collect indoor aerosol samples to monitor the effects of human breathing. In general, the instrumental options of aerosol collection have not been standardized yet and vary widely but the general principle of how they operate remains unchanged.

METAGENOMIC SEQUENCING

The current research into the taxonomic diversity of the microbiota in bioaerosols relies on the approaches employing next-generation sequencing. Total DNA is extracted from the total pool of microorganisms trapped in a filter or a liquid medium and further used in a metagenomic analysis. As such methods are being developed, it has become possible to identify the unculturable microorganisms that are the major component of the aerial biome [40]. The metagenomic findings obtained thus far have shown that the dominant species of microorganisms identified using this technique differ from those identified by conventional culturing methods [41], since > 99% of the microorganisms detected in the air cannot be grown under laboratory conditions [26]. The term "microbiome" has been coined and has become widely used; the following definition was provided in the *Microbiome* journal: "This term refers to the entire habitat, including the microorganisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions" [42]. However, in publications using the term "microbiome" in their titles or keywords, there is often a mismatch with this definition, since most studies focus on a single group (viruses, bacteria, fungi, or plants), or, in the best-case scenario, on a combination of two groups. Without going deep into the reasons for this state of affairs, in this review we only mention this fact and emphasize that when further using the term "microbiome," we mean the bacterial or fungal components of the microbiome or their combination, in line with the authors of the cited studies.

Hence, one can find a large number of publications related to research into the microbiomes in all types of natural objects, such as hot springs, lakes, seas, soil, the endogenous microbiota of organisms, etc. [43–46]; however, catastrophically few publications have focused on the metagenomic analysis of bioaerosols [47–52].

Metagenomic sequencing can be conveniently divided into two global directions: whole-genome metagenomic analysis and targeted sequencing (metabarcoding). In the former case, the entire DNA isolated from the sample is read, which allows one to talk about the taxonomic diversity, while on the other hand offering an opportunity to analyze its functional properties. However, the cost of the metagenomic approach is surely higher [48, 53] than that of metabarcoding, which is based on analyzing the highly conserved marker genes such as 16S (bacteria and archaea), ITS (fungi and plants), rbcL (plants), 18S (various eukaryotes), etc. [54–55]. Meanwhile, efficiency in taxonomic identification depends directly on the number of verified sequences in the specialized databases being used. Today, the most comprehensive databases are those for prokaryotes (16S) and fungi (ITS).

The exceptionally low content of microorganisms in the air, along with the significant variation in the composition of microbial ensembles, poses a serious problem in analyzing the biodiversity, the function spectrum, and metabolic activity of bioaerosol microbiota [56]. We would like to emphasize that studies of this type are a fundamental basis for identifying aspects of human–nature interactions, and in particular those related to the routes of disease transmission and potential impact on human health [57]. Nevertheless, only sporadic results of the metagenomic analysis of bioaerosols have been reported in Russia [58].

Bacterial microbiome in bioaerosols

In the near-surface layers of the atmosphere, bacteria constitute a significant portion of bioaerosols: for example, in the Colorado mountains (USA), the average bacteria content among aerosol particles sized > 0.5 μm was 22% [47].

Aerosol bacteria can have a significant impact on the atmospheric chemistry, thus affecting human health [15]. For example, high air pollution levels can greatly alter the structure of the bacterial microbiome in humans [59]. On foggy days in Beijing, the contents of pathogenic *Halomonas* and *Shewanella* bacteria were found to increase [60], especially in autumn and early winter.

Metagenomic sequencing has revealed that the bacterial microbiome in bioaerosols is substantially biodiverse [61]. For example, 38 bacterial taxa were identified in the near-surface layers of the troposphere in urban areas [41]. Most studies demonstrated that Proteobacteria, Firmicutes, and Actinobacteria are the major phyla in the bacterial microbiome of the lower [41, 62–63] and upper troposphere [50, 64–65]. Meanwhile, in the lower troposphere in urban areas, Firmicutes can make a significant contribution (20–30%), while such phyla as Cyanobacteria, Bacteroidetes, Chloroflexi, Acidobacteria, and Deinococcus-Thermus are the minor phyla (1–5% of the relative content of nucleotide sequences). However, other studies showed a high proportion of the Bacteroidetes phylum members in bioaerosols in the air above Japan after dust storms in Asia [17, 66], as well as in the air above eastern Australia [67]. The composition of the bacterial microbiome in the upper troposphere above the Noto Peninsula in Japan was quite specific, where it was demonstrated (although using fluorescence *in situ* hybridization) that 80% of all eubacteria on mineral aerosol particles were represented by *Bacillus subtilis* belonging to the phylum Firmicutes [68].

Various weather events have a significant impact on the composition and structure of the bacterial microbiome of bioaerosols. For example, the long-distance transport of dust particles aerosolized during dust storms by air currents over seas and continents is an important mechanism for the introduction of various microorganisms into local ecosystems [69]. Thus, storms in the Sahara Desert cause the penetration of dust particles into the atmosphere, which are then transported to Europe together with air masses; in particular, this leads to their accumulation in the Alpine snow at an altitude of > 3,000 m above sea level [70]. Bioindicators of dust particles transferred from Algeria were members of the phyla Gemmatimonadetes and Deinococcus-Thermus [70], which are known to occur in dry oligotrophic habitats with relatively high levels of solar radiation; it allows them to survive during the transfer, while maintaining their metabolic activity. Very small quantities of pathogenic bacteria can be transferred with dust particles over very long distances [70]. The human body surface is a more plausible (compared to other biotopes) source of pathogenic bacteria in the air [71]. It was revealed that there is a clear dependence between the structure and composition of the bacterial microbiome at an altitude of 10 m above ground level (an island and a peninsula in East Asia) and dust storms in Central Asia [69]. Meanwhile, dust particles act as ice nucleation centers [72]. Precipitation is another important mechanism of transferring microorganisms from the upper to the lower troposphere, as well as to the terrestrial surface [73]. This study has shown that the composition of the bacterial microbiome in precipitation (a) corresponded to the bioaerosol sources along the transfer route and (b) exhibited an obvious seasonal dynamics when the relative abundance of prevailing Proteobacteria decreased from summer to winter.

It is noteworthy that, as opposed to the mycobiome, whose indoor composition depended on its outdoor composition and was independent of people's activity indoors, the indoor biodiversity of the bacterial microbiome was dependent both on the outdoor bacterial microbiome [65] and on people's activity indoors [41]. However, outdoor air pollution may not affect the biodiversity of bacterial and archaeal ensembles in indoor bioaerosols, as it was shown in a study conducted in Beijing [74]. This indicates that there are different mechanisms of formation and dynamics of different microbiome components, which should be borne in mind when planning observational experiments.

Cyanobacteria, which cause various health problems when they are inhaled, may contribute substantially to the total load with airborne particles [75]. Picocyanobacteria were recently detected in the near-surface atmospheric layers above land or water bodies in Greenland and Antarctica [76], where soil and water aerosolization is the leading mechanism of aerosol formation. Their transfer by wind is considered to be the main source of Cyanobacteria in air.

A meta-analysis of the results of 42 studies, covering more than 3,000 bioaerosol samples, revealed increased bacterial diversity, and relative abundance of pathogens in the samples associated with anthropogenic activity at collection sites [71].

Mycobiome of bioaerosols

Aerosol mycobiomes vary greatly; however, at the phylum level, Basidiomycota and Ascomycota are the major components of the mycobiome in both the near-surface and higher troposphere layers (they can switch places in terms of dominance). Thus, the members of the phylum Ascomycota were dominant (more than two-thirds) in the near-surface air layer in the Colorado mountains at an altitude of > 3,000 m above sea level [77], as well as in the near-surface air layers in Kuwait at a significantly lower altitude [78]. Other researchers, however, revealed that the phylum Basidiomycota was dominant (⩾ 60%) [41, 63, 79], while the phylum Ascomycota accounted for about one-third of the fungal sequences. Interestingly, the proportion of members of the phylum Ascomycota (*Cladosporium* and *Alternaria*) resistant to atmospheric stress increased with altitude (500–800 m vs. 5–10 m) over the Gobi and the Taklimakan Deserts [80], which are the key suppliers of dust particles to the Asian atmosphere. In the near-surface air above a 3,043-m high mountain in Austria, members of the phylum Basidiomycota (Agaricomycetes) were dominant, followed by members of the phylum Ascomycota such as Dothideomycetes, Saccharomycetes, Sordariomycetes, Leotiomycetes, and Eurotiomycetes [64]. Ascomycetes, members of the family Davidiellaceae, accounted for 25% of the mycobiome in the direction from northeastern China towards Japan [81]. However, a recent study addressing fungal biodiversity in aerosols over Antarctica detected no members of this family among the dominant families of the mycobiome [82]. The fungus *Alternaria,* belonging to Pleosporaceae/Pleosporales/Dothideomycetes/Ascomycota, is often identified among the major dominant fungal species of surface air layers both in urban areas (Nanjing, Beijing, and Seoul) and under natural conditions (the desert in Kuwait) [41, 78, 83]. The cultivated fungal genera *Alternaria*, *Aspergillus*, *Penicillium*, *Cladosporium*, etc., which are well-known as the major components of aerosol mycobiota [84], may account for ⩽ 12% of the total number of marker nucleotide sequences in the metagenomic approach [41]. It should be borne in mind, however, that the relative content of *Alternaria* in the air can vary greatly (from 10 to 40%) over both rural and urban areas depending on the year [85]. A relationship between the mycobiome composition of near-surface aerosols and the vegetation type and condition (humidity of leaves) was revealed in the same study. Some papers describe a quite unexpected mycobiome composition (i.e., the one significantly differing from the data reported in other studies). Thus, the sequences of the genus *Candida* (Saccharomycetales/Saccharomycetes/Ascomycota) were shown to account for 54% of the mycobiome of the lower troposphere [81]. As for the near-surface layer, it was revealed in the same study [81] that the mycobiome consisted exclusively of *Aspergillus spp.* (Aspergillus/Aspergillaceae/Eurotiales/Eurotiomycetes/Ascomycota). It is obvious that the composition of indoor bioaerosols largely depends on that of the near-surface atmospheric outdoor air, being especially true for the mycobiome whose composition depended on that of outdoor bioaerosols and was virtually independent of human activity, as has been demonstrated in a study of indoor air in kindergartens conducted in Korea [41]. The diversity of the indoor mycobiome may depend on outdoor air pollution, as was shown in a study conducted in Beijing [74]. Similar to the bacterial microbiome, the mycobiome composition may vary depending on particular meteorological events: for example, the content of fungi belonging to the class Agaricomycetes/Basidiomycota [86], which release vast quantities of spores into the atmosphere after rains, increased significantly after a rain over the arid area of the Mediterranean.

CONCLUSIONS

Hence, the bioaerosol microbiome is a highly dynamic system. Variation in the microbiome composition and structure depends on a vast array of factors. Many of them mediate, disguise, or interfere with each other, thus preventing one from identifying unambiguous spatial and temporal regularities. Transfer of microorganisms over long distances by air currents in the upper troposphere has a crucial impact on the composition of the lower layers that humans are directly in contact with. This can be of great importance in terms of the transmission routes of certain diseases and the potential effect on human health, especially in the context of world population growth and environmental pollution. Therefore, the pressing need for strengthening Russia's position in terms of research and monitoring of airspace (and the microbiological components of bioaerosols in particular) cannot be overestimated. ●

This work was supported by the Russian Foundation for Basic Research (project No. 19-05-50032).

REFERENCES

1. Jaenicke R. // Science 2005. V. 308. P. 73.
2. Moelling K., Broecker F. // J. Environ. Public. Health. 2020. Art.1646943.
3. Jia Y., Chen Y., Yan P., Huang Q. // Aerosol Air Qual. Res. 2021. V. 21. Art. 200497.
4. Naqvi H.R., Datta M., Mutreja G., Siddiqui M.A., Naqvi D.F., Naqvi A.R. // Environ. Pollut. 2021. V. 268. Art. 115691.
5. Després V.R., Huffman A.J., Burrows S.M., Hoose C., Safatov A.S., Buryak G., Fröhlich-Nowoisky J., Elbert W., Andreae M.O., Pöschl U., et al. // Tellus Ser. B Chem. Phys. Meteorol. 2012. V. 64. Art.15598.
6. Douwes J., Thorne P., Pearce N., Heederik D. // Ann. Occup. Hyg. 2003. V. 47. P. 187–200.
7. Peccia J., Hernandez M. // Atmos. Environ. 2006. V. 40. P. 3941–3961.
8. Šantl-Temkiv T., Sikoparija B., Maki T., Carotenuto F., Amato P., Yao M., Morris C.E., Schnell R., Jaenicke R., Pöhlker C., et al. // Aerosol Sci. Technol. 2020. V. 54. P. 520–546.
9. Matthias-Maser S., Jaenicke R. // Atmos. Environ. 2000. V. 34. P. 3805–3811.
10. Morris C.E., Sands D.C., Bardin M., Jaenicke R., Vogel B., Leyronas C., Ariya P.A., Psenner R. // Biogeosci. Discuss. 2008. V. 5. P. 191–212.
11. Brodie E.L., DeSantis T.Z., Parker J.P.M., Zubietta I.X., Piceno Y.M., Andersem G.L. // Proc. Natl. Acad. Sci. USA. 2007. V. 104. P. 299–304.
12. Xie W., Li Y., Bai W., Hou J., Ma T., Zeng X., Zhang L., An T. // Front. Environ. Sci. Eng. 2021. V. 15. Art. 44.
13. Puspitasari F., Maki T., Shi G., Chen B., Kobayashi F., Hasegawa H., Iwasaka Y. // Air Quality, Atmosphere & Health. 2015. V. 9. P. 631–644.
14. Safatov A., Agafonov A., Arshinov M., Baklanov A., Belan B., Buryak G., Fofonov A., Generalov M., Kozlov A., Lapteva N., et al. // Atmospheric and Oceanic Optics. 2018. V. 31. P. 519–531.
15. Gong J., Qi J., E B., Yin Y., Gao D. // Environ. Pollut. 2020. V. 257. P. 113485.
16. Li M., Qi J., Zhang H., Huang S., Li L., Gao D. // Sci. Total Environ. 2011. V. 409. P. 3812–3819.
17. Park J., Tomoaki I., Masao N., Yamaguchi N. // Sci. Repts. 2016. V. 6. Art. 35706.
18. Janyasuthiwong S., Rungratanaubon T., Saiohai T. // Int. J. Sci. Innov. Technol. 2021. V. 4. P. 41–49.
19. Brągoszewska E., Mainka A., Pastuszka J.S. // Atmosphere. 2017. V. 8. Art. 239.
20. Shaffer B.T., Lighthart B. // Microb. Ecol. 1997. V. 34. P. 167–177.
21. Dong L., Qi J., Shao C., Zhong X., Gao D., Wan Cao W., Gao J., Bai R., Long G., Chu G. // Sci. Total Environ. 2016. V. 541. P. 1011–1018.
22. Fahlgren C., Bratbak G., Sandaa R.-A., Thyrhaug R., Zweifel U.L. // Aerobiologia. 2011. V. 27. P. 107–120.
23. Andreeva I.S., Safatov A.S., Puchkova L.I., Emelyanova E.K., Buryak G.A., Ternovoi V.A. // Optika Atmosfery i Okeana. 2021. V. 34. P. 408–413.
24. Safatov A.S., Andreeva I.S., Buryak G.A., Olkin S.E., Reznikova I.K., Belan B.D., Panchenko M.V., Simonenkov D.V. // Atmosphere. 2022. V. 13. P. 651.
25. Andreeva I.S., Safatov A.S., Puchkova L.I., Emelyanova E.K., Buryak G.A., Olkin S.E., Reznikova I.K., Ohlopkova O.V. // Bulletin of Nizhnevartovsk State University. 2019. № 2. P. 3–11.
26. Shammi M., Rahman M.M., Tareq S.M. // Front. Environ. Sci. 2021. V. 9. Art. 328.
27. Georgakopoulos D.G., Després V., Frohlich-Nowoisky J., Psenner R., Ariya P.A., Pósfai M., Ahern H.E., Moffett B.F., Hill T.C.J. // Biogeosciences. 2009. V. 6. P. 721–737.
28. Peccia J., Milton D.K., Reponen T., Hill J. // Environ. Sci. Technol. 2008. V. 42. P. 4631–4637.
29. Deguillaume L., Leriche M., Amato P., Ariya P. A., Delort A.-M., Pöschl U., Chaumerliac N., Bauer H., Flossmann A.I., Morris C.E. // Biogeosci. Discuss. 2008. V. 5. P. 841–870.
30. Christner B.C., Morris C.E., Foreman C.M., Cai R., Sands D.C. // Science. 2008. V. 319. P. 1214.
31. Amato P., Menager M., Sanseime M., Laj P., Mailhot G., Delort A.M. // Atmos. Environ. 2005. V. 39. P. 4143–4153.
32. Pratt K.A., DeMott P., French J., Wang Z., Westphal D.L., Heymsfield A.J., Twohy C.H., Prenni A.J., Prather K.A. // Nat. Geosci. 2009. V. 2. P. 398–401.
33. Yoo K., Lee T.K., Choi E.J., Yang J., Shukla S.K., Hwang S.I., Park J. // J. Environ. Sci. 2017. V. 51. P. 234–247.
34. Lierl M.B., Hornung R.W. // Ann. Allergy Asthma Immunol. 2003. V. 90. P. 28–33.
35. Dales R.E., Cakmak S., Burnett R.T., Judek S., Coates F., Brook J.R. // Am. J. Respir. Crit. Care Med. 2000. V. 162. P. 2087–2090.
36. Rosas I., Salinas E., Yela A., Calva E., Eslava C., Cravioto A. // Appl. Environ. Microbiol. 1997. V. 63. P. 4093–4095.
37. Henningson E.W., Ahlberg M.S. // J. Aerosol Sci. 1994. V. 25. P. 1459–1492.
38. Su X., Sutarlie L., Loh X.J. // Chem. Asian. J. 2020. V. 15. P. 4241–4255.
39. Lim J.H., Nam S.H., Kim J., Kim N.H., Park G.S., Maeng J.S., Yook S.J. // J. Biomech. Eng. 2022. V. 144. № 7. P. 071008. doi: 10.1115/1.4053504.
40. Garrido-Cardenas J.A., Manzano-Agugliaro F. // Curr. Genet. 2017. V. 63. P. 819–829.
41. Shin S.K., Kim J., Ha S.M., Oh H.S., Chun J., Sohn J., Yi H. // PLoS One. 2015. V. 10. Art. e0126960.
42. Marchesi J.R., Ravel J. // Microbiome. 2015. V. 3. Art. 31.
43. Hou J., Sievert S.M., Wang Y., Seewald J.S., Natarajan V.P., Wang F., Xiao X. // Microbiome. 2020. V. 8. Art. 102.
44. Osborne P., Hall L.J., Kronfeld-Schor N., Thybert D., Haerty W. // Environmental Microbiome. 2020. V. 15. Art. 20.
45. Bashir A.K., Wink L., Duller S., Schwendner P., Cockell C., Rettberg P., Mahnert A., Beblo-Vranesevic K., Bohmeier M., Rabbow E., et al. / / Microbiome. 2021. V. 9. Art. 50.
46. Zhou X., Leite M.F.A., Zhang Z., Tian L., Chang J., Ma L., Li X., van Veen J.A., Tian C., Kuramae E.E. // Environmental. Microbiome. 2021. V. 16. Art. 4.
47. Bowers R.M., McCubbin I.B., Hallar A.G., Fierer N. // Atmos. Environ. 2012. V. 50. P. 41–49.
48. Bowers R.M., Clements N., Emerson J.B., Wiedinmayer C., Hannigan M.P., Fierer N. // Environ. Sci. Technol. 2013. V. 47. P. 12097–12106.
49. Bertolini V., Gandolfi I., Ambrosini R., Bestetti G., Innocente E., Rampazzo G., Franzetti A. // Appl. Microbiol. Biotechnol. 2013. V. 97. P. 6561–6570.
50. DeLeon-Rodriguez N., Lathem T.L., Rodriguez-R L.M., Barazesh J.M., Anderson B.E., Beyersdorf A.J., Ziemba L.D., Bergin M., Nenes A., Konstantinidis K.T. // 2013. PNAS. V. 110. P. 2575–2580.
51. Serrano-Silva N., Calderon-Ezquerro M.C. // Environ.

Pollut. 2018. V. 235. P. 20–29
52. Mu F., Li Y., Lu R., Qi Y., Xie W., Bai W. // Atmosph. Res. 2020. V. 231. P. 104676.
53. Cao C., Jiang W., Wang B., Fang J., Lang J., Tian G., Jiang J., Zhu T. // Environ. Sci. Technol. 2014. V. 48. P. 1499–1507.
54. Xu J. // Genome. 2016. V. 59. P. 913–932.
55. Deiner K., Bik H.M., Machler E., Seymour M., Lacoursiere-Roussel A., Altermatt F., Creer S., Bista I., Lodge D.M., de Vere N., et al. // Mol. Ecol. 2017. V. 26. P. 5872-5895.
56. Luhung I., Uchida A., Lim S.B.Y., Gaultier N.E., KeeC., Lau K. J. X., Gusareva E.S., Heinle C.E., Wong A., Balakrishnan N. V., et al. npj Biofilms Microbiomes. 2021. V. 7. Art. 37.
57. Wang Z., Li J., Qian L., Liu L., Qian J., Lu B., Guo Z. // J. Vis. Exp. 2019. V. 143. Art. e58795.
58. Gusareva E.S., Gaultier N.P.E., Premkrishnan B.N.V., Kee C., Lim S. B.Y., Heinle C. E., Purbojati R.W., Nee A.P., Lohar S.R., Yanqing K., et al. // Sci. Rep. 2020. V. 10. P. 21515.
59. Fan X.-Y., Gao J.-F., Pan K.-L., Li D.-C., Dai H.-H., Li X. // Environ. Pollut. 2019. V. 251. P. 668-680.
60. Li W., Yang J., Zhang D., Li B., Wang E., Yuan H. // Front. Microbiol. 2018. V. 9. Art. 1741.
61. Ruiz-Gil T., Acuña J. J., Fujiyoshi S.,Tanaka D., Noda J., Maruyama F., Jorquera M.A. // Environ. Int. 2020. V. 145. Art. 106156.
62. Tang K., Huang, Z., Huang, J., Maki T., Zhang Sh., Ma X., Shi J., Jianrong B., Zhou T., Wang G., et al. // Atmospheric Chemistry and Physics Discussions. 2017. P. 1–41.
63. Pollegioni P., Mattioni C., Ristorini M., Occhiuto D., Canepari S., Korneykova M.V., Gavrichkova O. // Atmosphere. 2022. V. 13. Art. 224.
64. Els N., Greilinger M., Reisecker M., Tignat-Perrier R., Baumann-Stanzer K., Kasper-Giebl A., Sattler B., Larose C. // Front. Microbiol. 2020. V. 11. P. 980.
65. González-Martín C., Pérez-González C.J., González-Toril E., Expósito F.J., Aguilera Á., Díaz J.P. // Front Microbiol. 2021. V. 12. Art. 732961.
66. Yamaguchi N., Park J., Kodama M., Ichijo T., Baba T., Nasu M. // Microb. Environ. 2014. V. 29. P. 82-88.
67. De Deckker P., Munday C.I., Brocks J., O'Loingsigh T., Allison G.E., Hope J., Norman M., Stuut J., Tapper N., Kaars S.V.D. // Aeolian Res. 2014. V. 15. P. 133–149.
68. Maki T., Kobayashi F., Yamada M., Hasegawa H., Iwasaka Y. // Aerobiologia. 2013. V. 29. P. 341–354.
69. Maki T., Lee K. C., Kawai K., Onishi K., Hong C. S., Kurosaki Y., Shinoda M., Kai K., Iwasaka Y., Archer S.D.J., et al. // J. Geophys.Res.: Atmospheres. 2019. V. 124. P. 5579–5588.
70. Meola M., Lazzaro A., Zeyer J. // Front. Microbiol. 2015. V. 6. Art. 1454.
71. Jiang X., Wang C., Guo J., Hou J., Guo X., Zhang H., Tan J., Li M., Li X., Zhu H. // Environ. Men. Sci. Technol. 2022. V. 56. P. 9891–9902.
72. Maki T., Furumoto Sh., Asahi Yu.,Lee K.,Watanab K., Aoki K.,Murakami M., Tajiri T., Hasegawa H., Mashio A.,Iwasaka Y. // Atmosph. Chem. Phys. 2018. V. 18. P. 8155–8171.
73. Hiraoka S., Miyahara M., Fujii K., Machiyama A., Iwasaki W. // Front. Microbiol. 2017. V. 8. Art. 1506.
74. Zhou F., Ni M., Zhen Y., Su Y., W. Y., Zhu T., Shen F. // J. Aerosol. Sci. 2021. V. 156. Art. 105798.
75. Genitsaris S., Kormas K.A., Moustaka-Gouni M. // Front Biosci. 2011. V. 3. P. 772–787.
76. Trout-Haney J.V., Heindel R.C., Virginia R. A. // Environ. Microbiol. Rep. 2020. V. 12. P. 296–305.
77. Bowers R.M., Lauber C.L., Wiedinmyer C., Hamady M., Hallar A.G., Fall R., Knight R., Fierer N. // Appl. Environ. Microbiol. 2009. V. 75. P. 5121–5130.
78. Al Salameen F., Habibi N., Uddin S., Al Mataqi K., Kumar V., Al Doaij B., Al Amad S., Al Ali E., Shirshikhar F. // PLoS One. 2020. V. 15. Art. e0241283.
79. Hanson B., Zhou Y., Bautista E.J., Urch B., Speck M., Silverman F., Muilenberg M., Phipatanakul W., Weinstock G., Sodergren E., Gold D. R., Sordillo J.E. // Environ. Sci. Process Impacts. 2016. V. 18. P. 713–724.
80. Maki T., Chen B., Kai K., Kawai K., Fujita K., Ohara K., Kobayashi F., Davaanyam E.,Noda J., Minamoto Y., Shi G., Hasegawa H., Iwasaka Y. // Atmosph. Environ. 2019a. V. 214. Art. 116848.
81. Rodó X., Curcoll R., Robinson M., Ballester J., Burns J.C., Cayan D.R., Lipkin W.I., Williams B.L., Couto-Rodriguez M., Nakamura Y., Uehara R., Tanimoto H., Morguí J.A. // Proc. Natl. Acad. Sci. USA. 2014. V. 111. P. 7952–7957.
82. Rosa L.H., Pinto O., Convey P., Carvalho-Silva M., Rosa C.A., Câmara P. // Microb. Ecol. 2021. V. 82. P. 165–172.
83. Yang T., Han Y.P., Li L., Liu J.X. // Huan Jing Ke Xue. 2019. V. 40. P. 1680–1687. [Article in Chinese]
84. Nageen Y., Asemoloye M.D., Põlme S., Wang X., Xu S., Ramteke P.W., Pecoraro L. // BMC Microbiol. 2021. V. 21. Art. 134.
85. Hanson M., Petch G.M., Ottosen T.-B., Skjøth C.A. // Sci. Total Environ. 2022. V. 830. Art. 154491
86. Tang K., Sánchez-Parra B., Yordanova P., Wehking J., Backes A. T., Pickersgill D. A., Maier S., Sciare J., Pöschl U., Weber B., Fröhlich-Nowoisky J. // Biogeosciences. 2022. V. 19. P. 71–91.

Bioinformatics-Structural Approach to the Search for New D-Amino Acid Oxidases

D. L. Atroshenko[1,2], D. I. Golovina[1], E. P. Sergeev[1], M. D. Shelomov[1], A. G. Elcheninov[2], I. V. Kublanov[1,2], T. A. Chubar[1], A. A. Pometun[1,2], S. S. Savin[1], V. I. Tishkov[1,2*]
[1]Department of Chemistry, Lomonosov Moscow State University, Moscow, 119991 Russia
[2]Federal Research Centre "Fundamentals of Biotechnology" of RAS, Moscow, 119071 Russia
[*]E-mail: vitishkov@gmail.com

Received August 22, 2022; in final form, September 29, 2022
DOI: 10.32607/actanaturae.11812

ABSTRACT **D-amino acid oxidase (DAAO, EC 1.2.1.2) plays an important role in the functioning of prokaryotes as well as of lower (yeast and fungi) and higher eukaryotes (mammals). DAAO genes have not yet been found in archaean genomes. D-amino acid oxidase is increasingly used in various fields, which requires the development of new variants of the enzyme with specific properties. However, even within one related group (bacteria, yeasts and fungi, mammals), DAAOs show very low homology between amino acid sequences. In particular, this fact is clearly observed in the case of DAAO from bacteria. The high variability in the primary structures of DAAO severely limits the search for new enzymes in known genomes. As a result, many (if not most) DAAO genes remain either unannotated or incorrectly annotated. We propose an approach that uses bioinformatic methods in combination with general 3D structure and active center structure analysis to confirm that the gene found encodes D-amino acid oxidase and to predict the possible type of its substrate specificity. Using a homology search, we obtained a set of candidate sequences, modelled the tertiary structure of the selected enzymes, and compared them with experimental and model structures of known DAAOs. The effectiveness of the proposed approach for discrimination of DAAOs and glycine oxidases is shown. Using this approach, new DAAO genes were found in the genomes of six strains of extremophilic bacteria, and for the first time in the world, one gene was identified in the genome of halophilic archaea. Preliminary experiments confirmed the predicted specificity of DAAO from *Natronosporangium hydrolyticum* ACPA39 with D-Leu and D-Phe.**

KEYWORDS **D-amino acid oxidase, primary structure, ternary structure, modelling, AlphaFold 2, glycine oxidases.**

ABBREVIATIONS **DAAO – D-amino acid oxidase; GOX – glycine oxidase.**

INTRODUCTION

Any cell is a highly complex open-type multienzyme system, and depending on the complexity and specific state of functioning of the organism, the same enzyme can perform different physiological roles. A good example is D-amino acid oxidase (DAAO, EC 1.4.3.3). In bacteria, yeast, and fungi, the main role of this enzyme is limited to the utilization of exogenous D-amino acids (primarily D-Ala) [1, 2]. In higher eukaryotes – vertebrates and especially in mammals, the main role of DAAO is to maintain a certain level of D-amino acids, which are regulators of many important processes, primarily nervous activity. For example, a decrease in the level of D-Ser in the cerebrospinal fluid due to increased DAAO activity is associated with schizophrenia [3, 4]. In Alzheimer's and Parkinson's diseases, increased levels of D-Ala are observed in nervous tissues [4, 5]. Therefore, the search for effective and specific inhibitors of human DAAO seems to be relevant. D-amino acid oxidase is also widely used in practice [6–9]. For example, DAAO from the yeast *Trigonopsis variabilis* is used in the two-enzyme biocatalytic process for the production of 7-aminocephalosporonic acid (7-ACA) from cephalosporin C [10, 11]. This process reduces the consumption of organic solvents by 400 times compared to the previously used purely chemical method. The production of 7-ACA, used as a synthone for the production of semisynthetic cephalosporins of various generations, reaches several thousand tons per year.

The practical application of the enzyme requires the use of a biocatalyst with certain properties. There is no universal enzyme in nature. Its activity and specificity are stipulated by the role it per-forms in nature. In most biotechnological processes, the substrates and reaction conditions differ from those in nature. Therefore, when developing a new process, method of analysis and in other cases for each one, the properties of the biocatalyst are adjusted to the requirements of the process. This is usually done by protein engineering methods. Obviously, the enzyme whose properties are the closest to the required ones seems to be the optimal starting object. For this purpose, genes are searched for in sequenced genomes, the number of which is constantly increasing. The genes of D-amino acid oxidases from yeast and fungi were cloned, and the properties of the enzymes were studied from only 7 sources: *Fusarium solani* (FsoDAAO) [12], *Trigonopsis variabilis* (TvaDAAO) [13], *Rhodosporidium toruloides* (formerly *Rhodotorula gracilis*) (RtoDAAO) [14], *Pichia pastoris* (PpaDAAO) [15], *Candida boidinii* (CboDAAO) [16], *Rasamsonia emersonii* strain YA (RemDAAO) [17], and *Ogataea parapolymorpha* DL-1 [18]. In the latter case, the genes of five different DAAOs (OpaDAAO1 – OpaDAAO5) and one D-aspartate oxidase (DASPO) were identified, cloned, and expressed in *E. coli*. In the case of bacteria, only three enzymes have been cloned and described thus far: from *Rubrobacter xylanophilus* (RxyDAAO), *Streptomyces coelicolor* (ScoDAAO), and *Arthrobacter protophormiae* (AprDAAO) [9]. Currently, there are no data on the presence of potential *daao* genes in the archaean genomes in the literature and databases. The main reason for this state of affairs in the research and application of bacterial DAAOs is the difficulty in finding enzyme genes in bacterial genomes. A total of just over 10 DAAO genes have been identified, and all of them have been found in the genomes of bacteria belonging to *Acinetobacteria* [7, 9]. The difficulty of the search is related to the fact that the amino acid sequences of DAAOs are highly variable. Therefore, the traditional widely used homology search is a very difficult task. In addition, there is a closely related enzyme, glycine oxidase (GOX), which very often appears when searching for DAAO by homology with known enzymes of this type.

The second important point in the search for new DAAOs is the selection of candidates with properties closest to those needed. Common DAAOs, with the exception of the highly specific D-aspartate oxidase, exhibit broad substrate specificity. Depending on the source, the spectrum of substrate specificity varies greatly, and activity with different D-amino acids may differ by an order of magnitude or more. Moreover, in some cases, the substrate specificity of DAAOs may have special requirements. For example, when developing methods for diagnosing neurodegenerative diseases, DAAOs that are active with D-Ser but not with D-Ala and vice versa are required [5]. Therefore, to select a DAAO with the desired substrate specificity (if its description is not available), we have to clone and express a representative set of enzymes, purify and study their catalytic properties and select the best one. Obviously, this procedure is laborious, time-consuming, and expensive.

We have proposed a bioinformatic-structural approach that allows us to show with high reliability the belonging of candidate enzymes exactly to DAAOs, discriminate them from glycine oxidases, and, using the correlation between substrate specificity and experimental or model structures of known DAAOs, make a reasonable assumption about the spectrum of substrate specificity. Particular emphasis is placed on using data on enzymes from the thermotolerant yeast *O. parapolymorpha* DL-1 (five OpaDAAO and OpaDASPO) because five of the six enzymes exhibit unusual dependencies of stability and activity on medium pH and have a very interesting and promising spectrum of substrate specificity. This approach has been successfully tested on a number of sequences from extremophilic bacteria. The presence of the *daao* gene in the genome of a halophilic archaea has been shown for the first time in the world.

EXPERIMENTAL

Bioinformatic search for potential DAAO genes

The homology search for new DAAOs was performed using BLASTp software (https://blast.ncbi.nlm.nih.gov/Blast.PAGE=Proteins) against a database of translated protein sequences from the genomes of extremophilic bacteria. The UniProt NCBI was used as the main source. We also searched the genomes of bacteria and archaea whose sequences were determined during work under Agreement No. 075-15-2021-1396 dated 10/26/2021 (Federal Research Program for the Development of Genetic Technologies for 2019–2027). The sequences that showed the highest homology were selected for further work.

Multiple alignment of the selected sequences and a number of known bacterial and yeast sequences was carried out with Clustal X 1.83 software.

Construction and analysis of DAAO model structures

The open-access online server for AlphaFold2 was used to build model structures of enzymes [19, 20]. MMseqs2 software was used for multiple alignments, and three cycles of prediction refinement were performed for each model. Five models were generated, and the best variant was chosen based on the pLDDT value [19]. All obtained structures had a pLDDT greater than 90. The FAD molecule was incorporated by optimizing the position in the globule and bond geometry with Coot software [21].

Substrate docking was performed using AutoDock software [22] with GPU acceleration [23]. The following parameters were used for docking: ga_pop_size = 150, ga_num_evals = 25000000, ga_run = 20, ga_mutation_rate 0.02–0.08, Solis-Wets method. Docking results were selected based on the positions of the carboxyl group, amino group, and Cα-atom of the D-amino acid suitable for catalysis of reaction. The corresponding positions were selected based on the crystal structures of RtoDAAO in complex with D-alanine/iminopyruvate (PDBID 1C0P) and pkDAAO (from pig kidney) in complex with iminotryptophan (PDBID 1DDO). The position of the D-amino acid side chain was chosen based on the potential interactions of the substrate with DAAO.

The RMSD between structures was calculated by Cα atoms using the "align" command of the PyMol software package (The PyMOL Molecular Graphics System, Version 2.1.0, Schrödinger, LLC). Five cycles of structural emission deviations (parameter "cycles") were used to calculate the RMSD.

The structures were also visualized using PyMol software (The PyMOL Molecular Graphics System, Version 2.1.0, Schrödinger, LLC).

RESULTS AND DISCUSSION

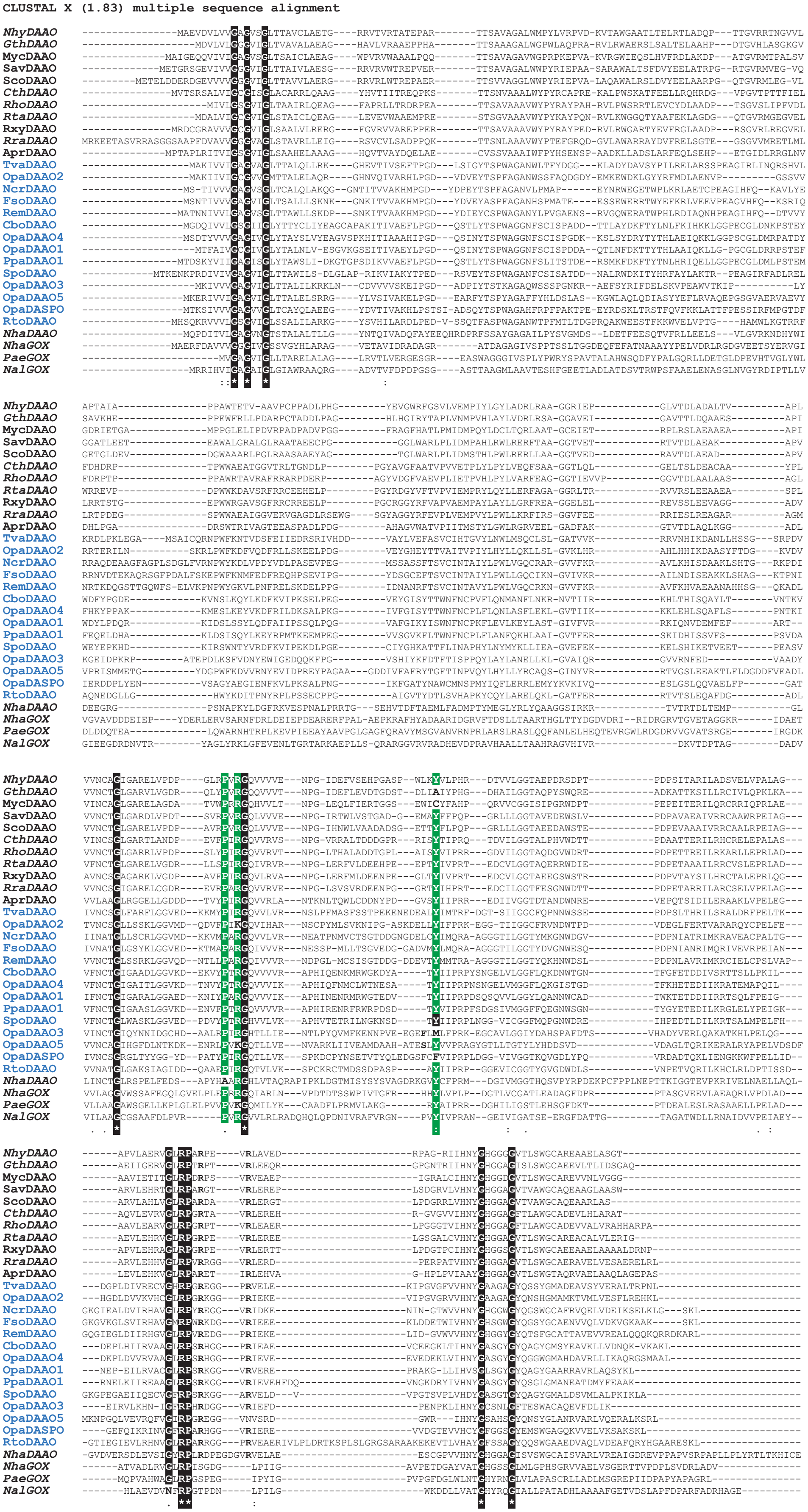
Homology search for new DAAOs from extremophilic bacteria and archaea

The search for new potential DAAOs was performed using the UniProt NCBI database for bacterial genomes and the joint database of sequenced genomes of extremophilic microorganisms of Lomonosov Moscow State University and Federal Research Centre "Fundamentals of Biotechnology" of the Russian Academy of Sciences (FRC Biotechnology RAS). Amino acid sequences of enzymes from the yeast *R. toruloides* (better known as *R. gracilis*), *T. variabilis*, *C. boidinii*, *O. parapolymorpha* DL-1 (five DAAOs and one DASPO) and the bacteria *A. protophormiae*, *R. xylanophilus* and *S. coelicolor* were used as references. Detailed information about the sources of DAAO sequences used in this work is presented in *Table. 1.* In the case of bacteria, sequences of only those enzymes with proven oxidase activity were used. First, the new enzymes were compared with the most well-studied DAOOs from *R. toruloides* and *T. variabilis*. Special attention was given to five DAAOs and one DASPO from the yeast *O. parapolymorpha* DL-1 because this is the only organism thus far in which so many paralogous enzymes have been obtained and studied. The *daao* and *daspo* genes from the yeast *O. parapolymorpha* DL-1 were cloned and expressed in *E. coli* cells in the active form. Four DAAOs and DASPO were obtained in a highly purified form, their catalytic parameters with D-amino acids were determined, their activity and stability dependencies were studied at different medium pH values, and their thermal stability at pH values optimal for stability was studied. The amino acid sequences of vertebrate DAAO were not used because they initially had low homology with microbial enzymes [1, 2, 9].

The search for DAAO homologues in bacterial genomes deposited at UniProt NCBI found a large number of candidate sequences, but the level of homology did not exceed 30%. An expert evaluation of the search results showed that the vast majority of sequences with a homology level of less than 23% cannot be attributed to DAAO. Therefore, only sequences from thermophilic bacteria with homology levels of 24–30% were selected for further work. Expert evaluation of these proteins based on conserved residues (see the next section) allowed us to narrow down the set to the sequences characteristic of DAAO and GOX. A similar sequence of procedures was used when searching for potential DAAO genes in the genomes of extremophiles and archaea in the database of Lomonosov Moscow State University and the FRC Biotechnology RAS.

Comparison of the amino acid sequences of the new DAAOs with known enzymes from bacteria, yeast, and fungi

Figure 1 shows some of the results of multiple alignment of the identified sequences (names of new enzymes are in bold italics) with the sequences of reference DAAOs. To avoid cluttering the already large figure, it does not show all the search results in the UniProt NCBI database. We left only the sequences of five DAAOs from thermomophilic microorganisms and did not provide data for glycine oxidases. However, multiple alignments of the glycine oxidase sequences were also performed, and model structures were constructed, based on the analysis of which they were assigned to GOX. Four sequences were selected from the genome databases of MSU and the FRC Biotechnology RAS after expert evaluation: one each from the bacteria *Natronosporangium hydrolyticum* ACPA39 [24] and *Natroglycomyces albus* ACPA22 [25] (this enzyme ended up being a glycine oxidase) and two from the archaea *Natrarchaeobius halalkaliphilus* AArcht4 [26]. In addition, sequences from two pathogens, *Mycobacterium tuberculosis* (MycDAAO) and *Pseudomonas aeruginosa* (PaeGOX), also found by



```
CLUSTAL X (1.83) multiple sequence alignment

NhyDAAO     ----------------MAEVDVLVVGAGVSGLTTAVCLAETG----RRVTVRTATEPAR-------TTSAVAGALWMPYLVRPVD-KVTAWGAATLTELRTLADQP---TTGVRRTNGVVL
GthDAAO     -------------------MDVLVLGGGVIGLTVAVALAEAG----HAVLVRAAEPPHA-------TTSAAAGALWGPWLAQPRA-RVLRWAERSLSALTELAAHP---DTGVHLASGKGV
MycDAAO     --------------MAIGEQQVIVIGAGVSGLTSAICLAEAG----WPVRVWAAALPQQ-------TTSAVAGAVWGPRPKEPVA-KVRGWIEQSLHVFRDLAKDP---ATGVRMTPALSV
SavDAAO     -------------METGRSGEVIVVGGGVIGLTTAIVLAESG----RRVRVWTREPVER-------TTSAVAGALWWPYRIEPAA-SARAWALTSFDVYEELATRPG--RTGVRMVEG-VQ
ScoDAAO     ---------METELDDERDGEVVVVGGGVIGLTTAVVLAERG----RRVRLWTREPAER-------TTSVVAGGLWWPYRIEPVA-LAQAWALRSLDVYEELAARPG--QTGVRMLEG-VL
CthDAAO     ---------------MVTSRSALVIGCGISGLACARRLQAAG----YHVTIITREQPKS-------TTSNVAAALWYPYRCAPRE-KALPWSKATFEELLRQHRDG---VPGVTPTTFIEL
RhoDAAO     ---------------------MIVLGSGVIGLTAAIRLQEAG----FAPRLLTRDRPEA-------TTSAVAAAVWYPYRAYPAH-RVLPWSRRTLEVCYDLAADP---TSGVSLIPFVDL
RtaDAAO     -------------------MDALVIGCGVIGLSTAICLQEAG----LEVEVWAAEMPRE-------STSGVAAAVWYPYKAYPQN-RVLKWGGQTYAAFEKLAGDG---QTGVRMGEGVEL
RxyDAAO     ---------------MRDCGRAVVVGCGVIGLSAALVLRERG----FGVRVVAREPPER-------TTSAVAAAVWYPYRAYPED-RVLRWGARTYEVFRGLAADP---RSGVRLREGVEL
RraDAAO     MRKEETASVRRASGGSAAPFDVAVVGGGVAGLSTAVRLLEIG----RSVCVLSADPPQK-------TTSNLAAAVWYPTEFGRQD-GVLAWARRAYDVFRELSGTE---GSGVVMRETLML
AprDAAO     --------------MPTAPLRITVIGSGVIGLSAAHELAAAG----HQVTVAYDQELAE-------CVSSVAAAIWFPYHSENSP-AADKLLADSLARFEQLSEHP---ETGIDLRRGLNV
TvaDAAO     ------------------MAKIVVIGAGVAGLTTALQLLRK----GHEVTIVSEFTPGD---LSIGYTSPWAGANWLTFYDGG---KLADYDAVSYPILRELARSSPEAGIRLINQRSHVL
OpaDAAO2    ------------------MAKIIVIGCGVVGMTTALELAQR----GHNVQIVARHLPGD---VDVEYTSPFAGANWSSFAQDGDY-EMKEWDKLGYYRFMDLAENVP---------GSSVV
NcrDAAO     -----------------MS-TIVVVGAGVSGLTCALQLAKQG--GNTITVVAKHMPGD---YDPEYTSPFAGANVLPMAP------EYNRWEGETWPLKRLAETCPEAGIHFQ--KAVLYE
FsoDAAO     -----------------MSNTIVVVGAGVIGLTSALLLSKNK---GNKITVVAKHMPGD---YDVEYASPFAGANHSPMATE----ESSEWERRTWYEFKRLVEEVPEAGVHFQ--KSRIQ
RemDAAO     ----------------MATNNIVVLGAGVSGLTTAWLLSKDP---SNKITVAAKHMPGD---YDIEYCSPWAGANYLPVGAENS--RVGQWERATWPHLRDIAQNHPEAGIHFQ--DTVVY
CboDAAO     -----------------MGDQIVVLGSGIIGLYTTYCLIYEAGCAPAKITIVAEFLPGD---QSTLYTSPWAGGNFSCISPADD--TTLAYDKFTYLNLFKIHKKLGGPECGLDNKPSTEY
OpaDAAO4    -----------------MSDTYVVVGAGIVGLYTAYSLVYEAGVSPKHITIAAAHIPGD---QSINYTSPWAGGNFSCISPGDK--KSLSYDRYTYTHLAEIQKKLGGPSCGLDMRPATDY
OpaDAAO1    -------------------MTFAIVGCGIVGLYTALNLV-ESGVKGSEITIVAEYLPGD---QSINYTSPWAGGNFSCISPDDA--QTLNFDKTTYTHLAAIQKLLG-PGCGLDRRPSTEF
PpaDAAO1    ----------------MTDSKYVIIGAGISGLYTAWSLI-DKGTGPSDIKVVAEFLPGD---QSTLYTSPWAGGNFSLITSTDE--RSMKFDKFTYTNLHRIQELLGGPECGLDMLPSTEM
SpoDAAO     ------------MTKENKPRDIVIVGAGVIGLTTAWILS-DLGLAP-RIKVIAKYTPED---RSVEYTSPWAGANFCSISATDD--NALRWDKITYHRFAYLAKTR--PEAGIRFADLREL
OpaDAAO3    -------------------MKIVVVGAGVIGLTTALILKRKLN---CDVVVVSKEIPGD---ADPIYTSTKAGAQWSSSPGNKR--AEFSYRIFDELSKVPEAWVTKIP----------LY
OpaDAAO5    -----------------MKERIVVIGAGVIGLTTALELSRRG----YLVSIVAKELPGD---EARFYTSPYAGAFFYHLDSLAS--KGWLAQLQDIASYYEFLRVAQEPGSGVAERVAEVY
OpaDASPO    -----------------MTKSIVVVGAGVVGLTCAYQLAEEG----YDVTIVAKHLPSTSI-ADSQYTSPWAGAHFRPFPAKTPE-EYRDSKLTRSTFQVFKKLATTFPESSIRFMPGTDF
RtoDAAO     ---------------MHSQKRVVVLGSGVIGLSSALILARKG----YSVHILARDLPED-V-SSQTFASPWAGANWTPFMTLTDGPRQAKWEESTFKKWVELVPTG----HAMWLKGTRRF
NhaDAAO     -----------------MQPDITVLGAGVNGTSTALALTLLG----YNTQIVADQFAYEEQHRDPRFSSAYGAGAILPYSVGMDS--LDETFEESQTVFRLLEELS---VLGVRKNDHYWI
NhaGOX      ---------------MAERFDAVVVGGGIVGSSVGYHLARAG----VETAVIDRADAGR-------ATDAGAGIVSPPTSSLTGGDEQFEFATNAAAYYPELVDRLRGDGVEETSYERVGI
PaeGOX      -----------------------MVGAGVIGLLTARELALAG----LRVTLVERGESGR----EASWAGGGIVSPLYPWRYSPAVTALAHWSQDFYPALGQRLLDETGLDPEVHTVGLYWL
NalGOX      ------------------MRRIHVIGAGAIGLGIAWRAAQRG----ADVTVFDPDPGSG----ASTTAAGMLAAVTESHFGEETLADLATDSVTRWPSFAELENASGLNVGYRDIPTLLV
                                   ::* *  *                   :

NhyDAAO     APTAIA---------------PPAWTETV-AAVPCPPADLPHG--------YEVGWRFGSVLVEMPIYLGYLADRLRAA-GGRIEP-----------GLVTDLADALTV---------APL
GthDAAO     SAVKHE---------------PPEWFRLLPDARPCTADDLPAG--------HLHGIRYTAPLVNMPVHLAYLVDRLRSA-GGAVEI-----------GAVTTLDQAAES---------API
MycDAAO     GDRIETGA-------------MPPGLELIPDVRPADPADVPGG--------FRAGFHATLPMIDMPQYLDCLTQRLAAT-GCEIET-----------RPLRSLAEAAEA---------API
SavDAAO     GGATLEET-------------EAWALGRALGLRAATAEECPG-----------GGLWARLPLIDMPAHLRWLRERFTAA-GGTVET-----------RTVTDLAEAK-----------APV
ScoDAAO     GETGLDEV-------------DGWAAARLPGLRAASAAEYAG-----------TGLWARLPLIDMSTHLPWLRERLLAA-GGTVED-----------RAVTDLAEAD-----------APV
CthDAAO     FDHDRP---------------TPWWAEATGGVTRLTGNDLP--------PGYAVGFAATVPVVETPLYLPYLVEQFSAA-GGTLQL-----------GELTSLDEACAA---------YPL
RhoDAAO     FDRPTP---------------PPAWRTAVRAFRRARPDERP--------AGYVDGFVAEVPLIETPVHLPYLVARFEAG-GGTIEVVP---------GGVTDLAALAAS---------AGL
RtaDAAO     WRREVP---------------DPWWKDAVSRFRRCEEHELP--------PGYRDGYVFTVPVIEMPRYLQYLLERFAGA-GGRLTR-----------RVVRSLEEAAEA---------SPL
RxyDAAO     LRRTSTG--------------EPWWRGAVSGFRRCRREELP--------PGCRGGYRFVAPVAEMPAYLAYLLGRFREA-GGELEL-----------REVSSLEEVAGG---------ADV
RraDAAO     LRTPDEG--------------SPWWAEAIGGVERVGAGDLRSEWG----SGYAGGYRFEVPLVEMPVYLPWLLKRFIRS-GGVFEE-----------RRIESLREAGAR---------AGM
AprDAAO     DHLPGA---------------DRSWTRIVAGTEEASPADLPDG--------AHAGVWATVPIITMSTYLGWLRGRVEEL-GADFAK-----------GTVTDLAQLKGG---------ADL
TvaDAAO     KRDLPKLEGA----MSAICQRNPWFKNTVDSFEIIEDRSRIVHDD------VAYLVEFASVCIHTGVYLNWLMSQCLSL-GATVVK-----------RRVNHIKDANLLHSSG---SRPDV
OpaDAAO2    RRTERILN----------SKRLPWFKDFVQDFRLLSKEELPDG--------VEYGHEYTTVAITVPIYLHYLLQKLLSL-GVKLHR-----------AHLHHIKDAASYFTDG----KVDV
NcrDAAO     RRAQDEAAGFAGPLSDGLFVRNPWYKDLVPDYVDLPASEVPEG--------MSSASSFTSVCINTAIYLPWLVGQCRAR-GVVFKR-----------AVLKHISDAAKLSHTG---RKPDI
FsoDAAO     RRNVDTEKAQRSGFPDALFSKEPWFKNMFEDFREQHPSEVIPG--------YDSGCEFTSVCINTAIYLPWLLGQCIKN-GVIVKR-----------AILNDISEAKKLSHAG---KTPNI
RemDAAO     NRTKDQGSTTGQWFS-ELVKPNPWYGKVLPNFRELSKDELPPG--------IDNANRFTSVCINTAVYLPWLVGQCRKN-GVVFKR-----------AVFKHVAEAANAHHSG---QKADL
CboDAAO     WDFYPGDE------------KVNSLKQYLKDFKVIPKSELPEG--------VEYGISYTTWNFNCPVFLQNMANFLNKR-NVTIIR-----------KHLTHISQAYLT-------VNTKV
OpaDAAO4    FHKYPPAK------------KMESLKEYVKDFRILDKSALPKG--------IVFGISYTTWNFNCPLFLQNLASFLEKL-GVTIIK-----------KKLEHLSQAFLS-------PNTKI
OpaDAAO1    WDYLPDQR------------KIDSLSSYLQDFAIIPSSQLPQG--------VAFGIKYISWNFNCPKFLEVLKEYLAST-GIVFVR-----------RKIQNVDEMFEF-------ART--
PpaDAAO1    FEQELDHA------------KLDSISQYLKEYRPMTKEEMPEG--------VVSGVKFLTWNFNCPLFLANFQKHLAAI-GVTFER-----------SKIDHISSVFS--------PSVDA
SpoDAAO     WEYEPKHD------------KIRSWNTYVRDFKVIPEKDLPGE--------CIYGHKATTFLINAPHYLNYMYKLLIEA-GVEFEK-----------KELSHIKETVEET------PEASV
OpaDAAO3    KGEIDPKRP--------ATEPDLKSFVDNYEWIGEDQQKFPG---------VSHIYKFDTFTISPPQYLAYLANELLKL-GVAIQR-----------GVVRNFED-----------VAADY
OpaDAAO5    VPRISMMETG---------YDGPWFKDVVRNYEVIDPREYPAGAA---GDDIVFAFRYTGFTINPVQYLHYLLYRCAQS-GINYVR-----------RTVGSLEEAKTLFLDGDDFVEADL
OpaDASPO    IERDDPLYEN----------VSAGYAEGIENFKVLPKSALPNG--------IKFGATYNAWCMNSPMYIQFLERRLEMYYKVKIVQ-----------ESLGSLQQVAELFP-------GAT
RtoDAAO     AQNEDGLLG-------------HWYKDITPNYRPLPSSECPPG---------AIGVTYDTLSVHAPKYCQYLARELQKL-GATFER-----------RTVTSLEQAFDG---------ADL
NhaDAAO     DEEGRG---------------PSNAPKYLDGFRKVESPNALPRRTG---SEHVTDFTAEMLFADMPTYMEGLYRLYQAAGGSIRKR------------TVTRTDLTEMP----------GL
NhaGOX      VGVAVDDDEIEP----YDERLERVSARNFDRLDEIEPDEARERFPAL-AEPKRAFHYADAARIDGRVFTDSLLTAARTHGLTTYDGDVDRI--RIDRGRVTGVETAGGKR------IDAET
PaeGOX      DLDDQTEA-------------LQWARNHTRPLKEVPIEEAYAAVPGLGAGFQRAVYMSGVANVRNPRLARSLRASLQQFANLELHEQTEVRGWLRDGDRVVGVATSRGE-------IRGDK
NalGOX      GIEEGDRDNVTR--------YAGLYRKLGFEVENLTGRTARKAEPLLS-QRARGGVRVRADHEVDPRAVHAALLTAAHRAGVHIVR--------------DKVTDPTAG--------DADV

NhyDAAO     VVNCAGIGARELVPDP----GLRPVRGQVVVVE---NPG-IDEFVSEHPGASP--WLKYVLPHR---DTVVLGGTAEPDRSDPT-----------PDPSITARILADSVELVPALAG---
GthDAAO     VVNCTGLGARVLVGDR----QLYPVRGQQVVVT---NPG-IDEFLEVDTGDST--DLIAIYPHG---DHAILGGTAQPYSWQRE-----------ADKATTKSILLRCIVLQPKLKA---
MycDAAO     VINCAGLGARELAGDA----TVWPRRGQHVVLT---NPG-LEQLFIERTGGS---EWICYFAHP---QRVVCGGISIPGRWDPT-----------PEPEITERILQRCRRIQPRLAE---
SavDAAO     VVNCTGLGARDLVPDT----SVRPVRGQLVVVE---NPG-IRTWLVSTGAD-G--EMAYFFPQP---GRLLLGGTAVEDEWSLV-----------PDPAVAEAIVRRCAAWRPEIAG---
ScoDAAO     VVNCTGLGARELVPDP----AVRPVRGQLVVVE---NPG-IHNWLVAADADSG--ETTYFLPQP---GRLLLGGTAEEDAWSTE-----------PDPEVAAAIVRRCAALRPEIAG---
CthDAAO     VINCSGLGARTLANDP----EVFPIRGQVVRVS---NPG-VRRALTDDDGPR---RISYTIPRQ---TDVILGGTALPHVWDTT-----------PDAATTERILRHCRELEPALAS---
RhoDAAO     VVNCTGLGARRLVPDP----SLYPIRGQVVRVT---NPG-LTHALADDTGPL---AISYVIPRR---GDVILGGTAQDGVWDRT-----------PDPETTREILRKARLLEPRLAD---
RtaDAAO     VFNCTGLGARELVGDR----LLSPIRGQIVRVR---NPG-LERFVLDEEHPE---EPTYIVPRT---EDCVLGGTAQERRWDIE-----------PDPETAAAILRRCVSLEPRLAD---
RxyDAAO     AVNCSGAGARKLVGDP----AVFPIRGQVLRVA---NPG-LERFMLDEENPE---GLTYIVPRT---EDCVLGGTAEEGSWSTR-----------PDPVTAYSILHRCTALEPRLQG---
RraDAAO     VVNCSGIGARELCGDR----EVRPARGQVVRVE---NPG-LSVSVRDEENPG---GRTYIHPRT---EDCILGGTFESGNWDTT-----------PDPETARRILARCSELVPELAG---
AprDAAO     VVLAAGLRGGELLGDDD---TVYPIRGQVVRLA---NTKNLTQWLCDDNYPD---GVSYIIPRR---EDIIVGGTDTANDWNRE-----------VEPQTSIDILERAAKLVPELEG---
TvaDAAO     IVNCSGLFARFLGGVED--KKMYPIRGQVVLVR---NSLPFMASFSSTPEKENEDEALYIMTRF-DGT-SIIGGCFQPNNWSSE-----------PDPSLTHRILSRALDRFPELTK---
OpaDAAO2    TVNCSGLLSSKLGGVMD--QDVFPIKGQVIHAR---NSCPYMLSVKNIPG-ASKDELLYIFPRK-EGG-TIIGGCFRVNDWTPD-----------VDEGLFERTVARARQYCPELFE---
NcrDAAO     IINATGLLSCRLGGVMD--KKVMPARGQVVLVR---NEATPNMVCTSGTDDGNGDELCYIMQRA-AGGGTILGGTYMKGNWDGV-----------PDPNIATRIMKRAVEACPALTG---
FsoDAAO     IVNATGLGSYKLGGVED--KTMAPARGQIVVVR---NESSP-MLLTSGVEDG-GADVMYLMQRA-AGGGTILGGTYDVGNWESQ-----------PDPNIANRIMQRIVEVRPEIAN---
RemDAAO     VVNCTGLSSRKLGGVQD--NTLLPARGQIVVVR---NDPGL-MCSISGTDDG-DDEVTYMMTRA-AGGGTILGGTYQKHNWDSL-----------PDPNLAVRIMKRCIELCPSLVAP--
CboDAAO     VFNCTGIGAADLGGVKD--EKVYPTRGQVVVVR---APHIQENKMRWGKDYA-----TYIIPRPYSNGELVLGGFLQKDNWTGN-----------TFGFETDDIVSRTTSLLPKIL----
OpaDAAO4    VFNCTGIGAITLGGVKD--TNVYPTRGQVVVIK---APHIQFNMCLWTNESA-----TYIIPRPNSNGELVMGGFLQKGISTGD-----------TFKHETEDIIKRATEMAPQIL----
OpaDAAO1    IFNCTGIGARALGGAED--KNIYPARGQVVVIK---APHINENRMRWGTEDV-----TYIIPRPDSQSQVVLGGYLQANNWCAD-----------TWKTETDDIIRRTSQLFPEIG----
PpaDAAO1    VFNCTGIGAASLGGVKD--ENVFPTRGQVVVVR---APHIRENRFRWRPDSD-----TYVIPRPFSDGSIVMGGFFQEGNWSGN-----------TYGYETEDILKRGLELYPEIGK---
SpoDAAO     VFNCTGLWASKLGGVED--PDVYPTRGHVVLVK---APHVTETRILNGKNSD-----TYIIPRPLNGG-VICGGFMQPGNWDRE-----------IHPEDTLDILKRTSALMPELFH---
OpaDAAO3    VINCTGIQYNNIDGCHD--AALRPIRGHTLLIE---NTLPYQVMFKENNPVE-EGEFLMLFPRK-EGCAVLGGIYDAHSPAFDTS---------VHADYVERLQAKATKHLPELQG----
OpaDAAO5    VVNCAGIHGFDLNTKDK--ENRLPVKGQTLLVE---NVARKLIIVEAMDAAH-ATESLYVVPRAGYGTLLTGTYLYHDDSVD-------------VDAGLTQRIKERALRYAPELVDSDF
OpaDASPO    IVNCSGRGLTYYGG-YD--PATYPIRGQTLLVK---SPKDCPYNSETVTYQLEDGSFCFVIPRPLDGG-VIVGGTKQVGDLYPQ-----------VRDADTQKLIENGKKWFPELLID--
RtoDAAO     VVNATGLGAKSIAGIDD--QAAEPIRGQTVLVK---SPCKRCTMDSSDPASP-----AYIIPRP--GGEVICGGTYGVGDWDLS-----------VNPETVQRILKHCLRLDPTISSD--
NhaDAAO     LINCTGLRSPELFEDS---APYHAARGHLVTAQRAPIPKLDGTMISYSYSVAGDRKGVYCFPRM---DGIVMGGTHQSVPYRPDEKPCFPPLNEPTTKIGGTEVPKRIVELNAELLAQL-
NhaGOX      VVLAGGVWSSAFEGQLGVELPLEPRRGQIARLN---VPDTDTSSWPIVTGFR----HHYLVPLP--DGTLVCGATLEADAGFDPR-----------PTASGVEEVLAEAQRLVPDLAD---
PaeGOX      VLLAAGAWSGELLKPLGLELPVVPVKGQMILYK---CAADFLPRMVLAKG-------RYAIPRR--DGHILIGSTLEHSGFDKT-----------PTDEALESLRASAAELLPELAD---
NalGOX      VILAAGCGSAAFDLPVR------PVRGVVLRLRADQHQLPQDNIVRAFVRGN----PVYIVPRAN--GEIVIGATSE-ERGFDATTG---------TAGATWDLLRNAIDVVPEIAEY--
             . .*                 .  *                               :           :  .                                      . :

NhyDAAO     ------APVLAERVGLRPARPE----VRLAVED----------------------RPAG-RIIHNYGHGGGGVTLSWGCAREAAELASGT------------------------------
GthDAAO     ------AEIIGERVGLRPTRPT----VRLEEQR----------------------GPGNTRIIHNYGHGGAGISLSWGCAEEVLTLIDSGAQ----------------------------
MycDAAO     ------AAVIETITGLRPDRPS----VRVEAEP-----------------------IGRALCIHNYGHGGDGVTLSWGCAREVVNLVGGG------------------------------
SavDAAO     ------ARVLEHRTGLRPARGT----VRLEREP----------------------LSDGRVLVHNYGHGGAGVTVAWGCAQEAAGLAASW------------------------------
ScoDAAO     ------ARVLAHLVGLRPARDA----VRLERGT----------------------LPDGRRLVHNYGHGGAGVTVAWGCAQEAARLAS--------------------------------
CthDAAO     ------AQVLEVRVGLRPGRTA----VRLEREH----------------------R-GVGVVIHNYGHGGAGFTLAWGCADEVLHLARAT------------------------------
RhoDAAO     ------AAVLEARVGLRPGRPT----VRLEAER----------------------LPGGGTVIHNYGHGGAGFTLAWGCADEVVALVRAHHARPA-------------------------
RtaDAAO     ------AEVLEHRVGLRPGRPE----VRLEREE----------------------LGSGALCVHNYGHGGAGVTLSWGCAREACALVLERIG----------------------------
RxyDAAO     ------APVLEHRAGLRPGRPE----VRLERTT----------------------LPDGTPCIHNYGHGGSGVTLSWGCAEEAAELAAAALDRNP-------------------------
RraDAAO     ------ARVLEHHVGLRPVRRGG---VRLERD------------------------PERPATVHNYGHGGAGVTLSWGCAERAVELVESAERELRL------------------------
AprDAAO     ------LEVLEHKVGLRPARET----IRLEHVA----------------------G-HPLPVIAAYGHGGAGVTLSWGTAQRVAELAAQLAGEPAS------------------------
TvaDAAO     ---DGPLDIVRECVGHRPGREGG---PRVELE----------------------KIPGVGFVVHNYGAAGAGYQSSYGMADEAVSYVERALTRPNL------------------------
OpaDAAO2    ---HGDLDVVKVHCGLRPGRKGG---PRIEKE----------------------VIPGVGRVVHNYGAAGAGYQNSHGMAMKTVMLVESFLREHKL------------------------
NcrDAAO     GKGIEALDVIRHAVGLRPYREGG---VRIDKE----------------------NIN-GTWVVHNYGHGGWGYQGSWGCAFRVQELVDEIKSELKLG---SKL-----------------
FsoDAAO     GKGVKGLSVIRHAVGMRPWRKDG---VRIEEE----------------------KLDDETWIVHNYGHSGWGYQGSYGCAENVVQLVDKVGKAAK-----SKL-----------------
RemDAAO     GQGIEGLDIIRHGVGLRPVREDG---PRIEKE----------------------LID-GVWVVHNYGHGGYGYQTSFGCATTAVEVVREALQQQKQRRDKARL-----------------
CboDAAO     ---DEPLHIIRVAAGLRPSRHGG---PRIEAE----------------------VCEEGKLTIHNYGASGYGYQAGYGMSYEAVKLLVDNQK-VKAKL----------------------
OpaDAAO4    ---DKPLDVVRVAAGLRPSRHGG---PRIEVE----------------------EVEDEKLVIHNYGASGYGYQGGWGMAHDAVRLLIKAQRGSMAAL----------------------
OpaDAAO1    ---NEP-EILRVACGLRPSRKGG---VRVERE----------------------PRAKG-LLIHVSGLSGYGYQAGYGAARRAVRLAQSYKL----------------------------
PpaDAAO1    ---RNELKIIREAAGLRPSRKGG---VRIEVEHFDQ------------------VNGKDRYIVHNYGASGYGYQSGLGMANEATDMYFEAAK----------------------------
SpoDAAO     GKGPEGAEIIQECVGFRPSRKGG---ARVELD---V------------------VPGTSVPLVHDYGASGTGYQAGYGMALDSVMLALPKIKLA--------------------------
OpaDAAO3    ----EIRVLKHN-IGFRPHRDGG---VRIEFD------------------------PENPKLIHNYGCSNLGFTESWACAQEVFDLIK------------------------------
OpaDAAO5    MKNPGQLVEVRQFVGIRPGREGG---VNVSRD------------------------GWR---IHNYGSAHSGYQNSYGLANRVARLVQERALKSRL------------------------
OpaDASPO    ----GEFQIKRINVGFRPARHGG---VRIERE-----------------------VVDGTEVVHCYGFGGSGYEMSWGAGQKVVELVKSAKSKL--------------------------
RtoDAAO     -GTIEGIEVLRHNVGLRPARRGG---PRVEAERIVLPLDRTKSPLSLGRGSARAAKEKEVTLVHAYGFSSAGYQQSWGAAEDVAQLVDEAFQRYHGAARESKL-----------------
NhaDAAO     -GVDVERSDLEVSIGYRPLRDPEGDGVRVELAE-----------------------ESCGPVVHNYGHGGAGISVSWGCAISVARLVREAIGDREVPPAPVSRPAPLLPLYRTLTKHICE
NhaGOX      ------ATVADLRVGLRPISGDG-----LPIIG---------------------AVPETDGAYVATGHGSSGLMLGPHSGRVVAELVSGERTTVPDPLSVDRLADV--------------
PaeGOX      ------MQPVAHWAGLRPGSPEG-----IPYIG---------------------PVPGFDGLWLNTGHYRNGLVLAPASCRLLADLMSGREPIIDPAPYAPAGRL---------------
NalGOX      -------HLAEVDVNFRPGTPDN-----LPILG------------------------WKDDLLVATGHYRQGIALLPATADHLAAAAFGETVDSLAPFDPARFADRHAGE----------
                          . **          :                                     *    *
```

Fig. 1. Alignment of the amino acid sequences of D-amino acid oxidases from yeast, bacteria and archaea (the names are shown in blue and black, respectively). See *Table 1* for the name correspondence. Novel DAAO sequences from bacteria analysed in this work are shown in bold italics. Letters on a green background in the alignment refer to residues that were previously considered as conserved for D-amino acid binding in the DAAO active site

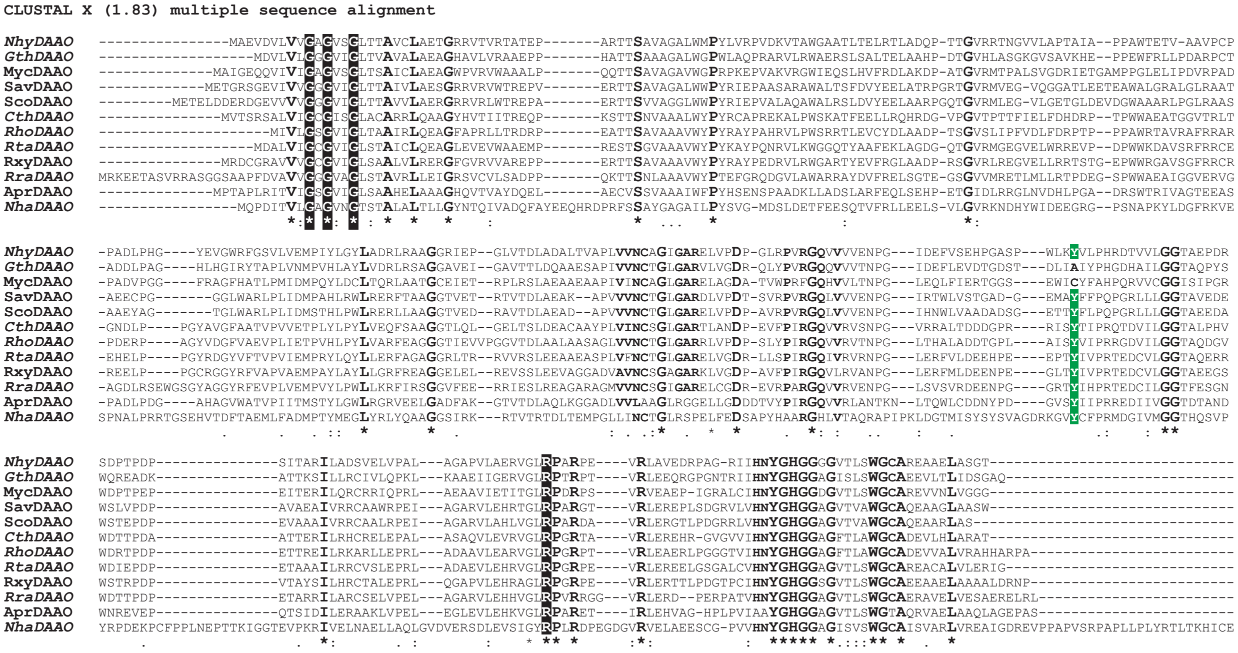
Table 1. D-amino acid oxidases and glycine oxidases and their sources*

No.	Short name	Source	Protein code in database
NCBI (GeneBank, UniProt)			
Bacteria			
1	***GthDAAO***	*Gandjariella thermophila*	WP_137812914.1
2	***CthDAAO***	*Chloracidobacterium thermophilum* B	WP_014099936.1
3	***RtaDAAO***	*Rubrobacter taiwanensis*	WP_132692836.1
4	***RraDAAO***	*Rubrobacter radiotolerans* DSM 5868	WP_084263988.1
5	***RbaDAAO***	*Rhodothermaceae bacterium* RA	ARA94025.1
6	RxyDAAO	*Rubrobacter xylanophilus*	BAP18969.1
7	MycDAAO	*Mycobacterium tuberculosis*	WP_003899072
8	SavDAAO	*Streptomyces avermitilis* MA-4680	BAC69383
9	ScoDAAO	*Streptomyces coelicolor* A3(2)	CAB40690
10	AprDAAO	*Arthrobacter protophormiae*	AY306197
11	***PaeGOX***	*Pseudomonas aeruginosa*	AAP81270
Fungi and yeasts			
12	TvaDAAO	*Trigonopsis variabilis*	AY514426
13	NcrDAAO	*Neurospora crassa*	EAA33029
14	FsoDAAO	*Fusarium solani*	BAA00692
15	RemDAAO	*Rasamsonia emersonii*	BBH51408
16	CboDAAO	*Candida boidinii*	BAB12222
17	PpaDAAO	*Pichia pastoris* CBS7435	SCV12162
18	SpoDAAO	*Schizosaccharomyces pombe*	NP_001342883
19	RtoDAAO	*Rhodosporidium toruloides* (*Rhodotorula gracilis*)	U60066
20	OpaDAAO1	*Ogataea parapolymorpha* DL-1	XP_013932717
21	OpaDAAO2	*Ogataea parapolymorpha* DL-1	XP_013937260
22	OpaDAAO3	*Ogataea parapolymorpha* DL-1	XP_013934816
23	OpaDAAO4	*Ogataea parapolymorpha* DL-1	XP_013937224
24	OpaDAAO5	*Ogataea parapolymorpha* DL-1	XP_013937169
25	OpaDASPO	*Ogataea parapolymorpha* DL-1	XP_013932178
The genome database of the Lomonosov Moscow University and the Federal Research Center of Biotechnology			
Bacteria			
26	***NhyDAAO***	*Natronosporangium hydrolyticum* ACPA39	lcl|CP070499.1_prot_QSB16697.1_2115
27	***NalGOX***	*Natroglycomyces albus* ACPA22	lcl|CP070496.1_prot_QSB06127.1_824
Archaea			
28	***NhaDAAO***	*Natrarchaeobius halalkaliphilus* AArcht4	2642575300
29	***NhaGOX***	*Natrarchaeobius halalkaliphilus* AArcht4	2642575587

*The new sequences of DAAOs from extremophilic microorganisms analysed in this study are shown in bold italics.

our homology search, are presented in the alignment. In the NCBI database, the *P. aeruginosa* protein is annotated as DAAO.

Multiple alignment of the selected sequences was performed using Clustal X 1.83 software (*Fig. 1*). This program is used because it itself builds a hierarchy in the homology of the given sequences. The results of this alignment gave quite expected results. As shown in *Fig. 1*, depending on the source, the enzymes are clearly divided into two groups: bacterial DAAOs are at the top, followed by yeast and fungal enzymes, immediately followed by DAAO from archaea, and then by glycine oxidases. The second interesting point is that the widely used and well-studied TvaDAAO has



```
CLUSTAL X (1.83) multiple sequence alignment

NhyDAAO    ----------------MAEVDVLVVGAGVSGLTTAVCLAETGRRVTVRTATEP-------ARTTSAVAGALWMPYLVRPVDKVTAWGAATLTELRTLADQP-TTGVRRTNGVVLAPTAIA--PPAWTETV-AAVPCP
GthDAAO    -------------------MDVLVLGGGVIGLTVAVALAEAGHAVLVRAAEPP-------HATTSAAAGALWGPWLAQPRARVLRWAERSLSALTELAAHP-DTGVHLASGKGVSAVKHE--PPEWFRLLPDARPCT
MycDAAO    --------------MAIGEQQVIVIGAGVSGLTSAICLAEAGWPVRVWAAALP-------QQTTSAVAGAVWGPRPKEPVAKVRGWIEQSLHVFRDLAKDP-ATGVRMTPALSVGDRIETGAMPPGLELIPDVRPAD
SavDAAO    -------------METGRSGEVIVVGGGVIGLTTAIVLAESGRRVRVWTREPV-------ERTTSAVAGALWWPYRIEPAASARAWALTSFDVYEELATRPGRTGVRMVEG-VQGGATLEETEAWALGRALGLRAAT
ScoDAAO    ---------METELDDERDGEVVVVGGGVIGLTTAVVLAERGRRVRLWTREPA-------ERTTSVVAGGLWWPYRIEPVALAQAWALRSLDVYEELAARPGQTGVRMLEG-VLGETGLDEVDGWAAARLPGLRAAS
CthDAAO    ---------------MVTSRSALVIGCGISGLACARRLQAAGYHVTIITREQP-------KSTTSNVAAALWYPYRCAPREKALPWSKATFEELLRQHRDG-VPGVTPTTFIELFDHDRP--TPWWAEATGGVTRLT
RhoDAAO    ---------------------MIVLGSGVIGLTAAIRLQEAGFAPRLLTRDRP-------EATTSAVAAAVWYPYRAYPAHRVLPWSRRTLEVCYDLAADP-TSGVSLIPFVDLFDRPTP--PPAWRTAVRAFRRAR
RtaDAAO    -------------------MDALVIGCGVIGLSTAICLQEAGLEVEVWAAEMP-------RESTSGVAAAVWYPYKAYPQNRVLKWGGQTYAAFEKLAGDG-QTGVRMGEGVELWRREVP--DPWWKDAVSRFRRCE
RxyDAAO    ---------------MRDCGRAVVVGCGVIGLSAALVLRERGFGVRVVAREPP-------ERTTSAVAAAVWYPYRAYPEDRVLRWGARTYEVFRGLAADP-RSGVRLREGVELLRRTSTG-EPWWRGAVSGFRRCR
RraDAAO    MRKEETASVRRASGGSAAPFDVAVVGGGVAGLSTAVRLLEIGRSVCVLSADPP-------QKTTSNLAAAVWYPTEFGRQDGVLAWARRAYDVFRELSGTE-GSGVVMRETLMLLRTPDEG-SPWWAEAIGGVERVG
AprDAAO    --------------MPTAPLRITVIGSGVIGLSAAHELAAAGHQVTVAYDQEL-------AECVSSVAAAIWFPYHSENSPAADKLLADSLARFEQLSEHP-ETGIDLRRGLNVDHLPGA--DRSWTRIVAGTEEAS
NhaDAAO    -----------------MQPDITVLGAGVNGTSTALALTLLGYNTQIVADQFAYEEQHRDPRFSSAYGAGAILPYSVG-MDSLDETFEESQTVFRLLEELS-VLGVRKNDHYWIDEEGRG--PSNAPKYLDGFRKVE
                                  *:* *: * : *  *   *    :                 *  ...   *               :              *:

NhyDAAO    -PADLPHG----YEVGWRFGSVLVEMPIYLGYLADRLRAAGGRIEP--GLVTDLADALTVAPLVVNCAGIGARELVPDP-GLRPVRGQVVVVENPG---IDEFVSEHPGASP---WLKYVLPHRDTVVLGGTAEPDR
GthDAAO    -ADDLPAG----HLHGIRYTAPLVNMPVHLAYLVDRLRSAGGAVEI--GAVTTLDQAAESAPIVVNCTGLGARVLVGDR-QLYPVRGQQVVVTNPG---IDEFLEVDTGDST---DLIAIYPHGDHAILGGTAQPYS
MycDAAO    -PADVPGG----FRAGFHATLPMIDMPQYLDCLTQRLAATGCEIET--RPLRSLAEAAEAAPIVINCAGLGARELAGDA-TVWPRFGQHVVLTNPG---LEQLFIERTGGS----EWICYFAHPQRVVCGGISIPGR
SavDAAO    -AEECPG-------GGLWARLPLIDMPAHLRWLRERFTAAGGTVET--RTVTDLAEAK--APVVVNCTGLGARDLVPDT-SVRPVRGQLVVVENPG---IRTWLVSTGAD-G---EMAYFFPQPGRLLLGGTAVEDE
ScoDAAO    -AAEYAG-------TGLWARLPLIDMSTHLPWLRERLLAAGGTVED--RAVTDLAEAD--APVVVNCTGLGARELVPDP-AVRPVRGQLVVVENPG---IHNWLVAADADSG---ETTYFLPQPGRLLLGGTAEEDA
CthDAAO    -GNDLP----PGYAVGFAATVPVVETPLYLPYLVEQFSAAGGTLQL--GELTSLDEACAAYPLVINCSGLGARTLANDP-EVFPIRGQVVRVSNPG---VRRALTDDDGPR----RISYTIPRQTDVILGGTALPHV
RhoDAAO    -PDERP----AGYVDGFVAEVPLIETPVHLPYLVARFEAGGGTIEVVPGGVTDLAALAASAGLVVNCTGLGARRLVPDP-SLYPIRGQVVRVTNPG---LTHALADDTGPL----AISYVIPRRGDVILGGTAQDGV
RtaDAAO    -EHELP----PGYRDGYVFTVPVIEMPRYLQYLLERFAGAGGRLTR--RVVRSLEEAAEASPLVFNCTGLGARELVGDR-LLSPIRGQIVRVRNPG---LERFVLDEEHPE----EPTYIVPRTEDCVLGGTAQERR
RxyDAAO    -REELP----PGCRGGYRFVAPVAEMPAYLAYLLGRFREAGGELEL--REVSSLEEVAGGADVAVNCSGAGARKLVGDP-AVFPIRGQVLRVANPG---LERFMLDEENPE----GLTYIVPRTEDCVLGGTAEEGS
RraDAAO    -AGDLRSEWGSGYAGGYRFEVPLVEMPVYLPWLLKRFIRSGGVFEE--RRIESLREAGARAGMVVNCSGIGARELCGDR-EVRPARGQVVRVENPG---LSVSVRDEENPG----GRTYIHPRTEDCILGGTFESGN
AprDAAO    -PADLPDG----AHAGVWATVPIITMSTYLGWLRGRVEELGADFAK--GTVTDLAQLKGGADLVVLAAGLRGGELLGDDDTVYPIRGQVVRLANTKN--LTQWLCDDNYPD----GVSYIIPRREDIIVGGTDTAND
NhaDAAO    SPNALPRRTGSEHVTDFTAEMLFADMPTYMEGLYRLYQAAGGSIRK---RTVTRTDLTEMPGLLINCTGLRSPELFEDSAPYHAARGHLVTAQRAPIPKLDGTMISYSYSVAGDRKGVYCFPRMDGIVMGGTHQSVP
                          .      .   . ::  *       *  .                  : . .:*  .  *  *     .  *: :   ..    :   .                 .:   :  **

NhyDAAO    SDPTPDP---------------SITARILADSVELVPAL---AGAPVLAERVGLRPARPE----VRLAVEDRPAG-RIIHNYGHGGGGVTLSWGCAREAAELASGT------------------------------
GthDAAO    WQREADK---------------ATTKSILLRCIVLQPKL---KAAEIIGERVGLRPTRPT----VRLEEQRGPGNTRIIHNYGHGGAGISLSWGCAEEVLTLIDSGAQ----------------------------
MycDAAO    WDPTPEP---------------EITERILQRCRRIQPRL---AEAAVIETITGLRPDRPS----VRVEAEP-IGRALCIHNYGHGGDGVTLSWGCAREVVNLVGGG------------------------------
SavDAAO    WSLVPDP---------------AVAEAIVRRCAAWRPEI---AGARVLEHRTGLRPARGT----VRLEREPLSDGRVLVHNYGHGGAGVTVAWGCAQEAAGLAASW------------------------------
ScoDAAO    WSTEPDP---------------EVAAAIVRRCAALRPEI---AGARVLAHLVGLRPARDA----VRLERGTLPDGRRLVHNYGHGGAGVTVAWGCAQEAARLAS--------------------------------
CthDAAO    WDTTPDA---------------ATTERILRHCRELEPAL---ASAQVLEVRVGLRPGRTA----VRLEREHR-GVGVVIHNYGHGGAGFTLAWGCADEVLHLARAT------------------------------
RhoDAAO    WDRTPDP---------------ETTREILRKARLLEPRL---ADAAVLEARVGLRPGRPT----VRLEAERLPGGGTVIHNYGHGGAGFTLAWGCADEVVALVRAHHARPA-------------------------
RtaDAAO    WDIEPDP---------------ETAAAILRRCVSLEPRL---ADAEVLEHRVGLRPGRPE----VRLEREELGSGALCVHNYGHGGAGVTLSWGCAREACALVLERIG----------------------------
RxyDAAO    WSTRPDP---------------VTAYSILHRCTALEPRL---QGAPVLEHRAGLRPGRPE----VRLERTTLPDGTPCIHNYGHGGSGVTLSWGCAEEAAELAAAALDRNP-------------------------
RraDAAO    WDTTPDP---------------ETARRILARCSELVPEL---AGARVLEHHVGLRPVRRGG---VRLERD--PERPATVHNYGHGGAGVTLSWGCAERAVELVESAERELRL------------------------
AprDAAO    WNREVEP---------------QTSIDILERAAKLVPEL---EGLEVLEHKVGLRPARET----IRLEHVAG-HPLPVIAAYGHGGAGVTLSWGTAQRVAELAAQLAGEPAS------------------------
NhaDAAO    YRPDEKPCFPPLNEPTTKIGGTEVPKRIVELNAELLAQLGVDVERSDLEVSIGYRPLRDPEGDGVRVELAEESCG-PVVHNYGHGGAGISVSWGCAISVARLVREAIGDREVPPAPVSRPAPLLPLYRTLTKHICE
                .                  .  *:       . :        :    * ** *      :*:           :  ***** *.:::** *  .  *
```

Fig. 2. Alignment of amino acid sequences of D-amino acid oxidases from bacteria and archaea after screening out glycine oxidase sequences. See *Table 1* for the name correspondence. New DAAO sequences from bacteria analysed in this work are shown in bold italics. The Tyr residue that was previously considered conserved for binding D-amino acids in the active site is shown with letters on a green background in the alignment

the highest homology with bacterial DAAOs, while the second, even more thoroughly studied RtoDAAO, is at the very end of the list before archaea.

As already mentioned, when searching for the genes of target enzymes in new sources, an approach based on the homology of proteins that perform the same function (e.g., catalyze the same reaction) is used. In some cases, such enzymes have very high homology in the substrate-binding and catalytic domains, and solving such a problem is not very difficult. A good example is NAD(P)$^+$-dependent formate dehydrogenase (FDH), which consists of two identical subunits and does not have cofactors in the active center. The degree of homology between FDHs even from evolutionarily distant sources (e.g., bacteria and higher plants) is at least 55%, and a large number of sufficiently extended (up to 10–15 amino acid residues) conserved sequences in all parts of the active center are observed in multiple alignments [27–29]. In DAAO, the level of homology does not exceed 30%, which is much lower. In this case, information on the conserved and catalytically important amino acid residues could help to annotate the gene. However, in the case of DAAO, this approach is of little use. A characteristic feature of the catalytic mechanism of FAD-containing enzymes is the transfer of the hydride ion from the substrate to the isoalkoxazine ring of the cofactor proceeds without significant participation of the amino acid residues of the enzyme, whose main role is to form the proper conformation of the active centre necessary for catalysis and the participation of a number of residues in the binding of FAD and D-amino acids. In the case of a cofactor, the fingerprint sequence GxGxxG must be present at the N-terminus of the enzyme [30]. The presence of arginine and tyrosine residues (R285 and Y223 in RtoDAAO and R302 and Y243 in TvaDAAO) in the active site was considered mandatory for binding the carboxyl group of the D-amino acid. Similar residues are also present in mammalian enzymes [1, 2]. However, the expansion of the set of compared sequences indicates that only the fingerprint sequence in the FAD-binding domain and the arginine residue participating in binding the carboxyl group remain conserved. Note that the mobility of this Arg residue is strongly restricted by the neighboring conserved proline residue (ArgPro pair, *Fig. 1*, fourth row of alignment). The tyrosine residue (*Fig. 1*, third row of alignment, middle) is not conserved – two of the six OpaDAAOs as well as two bacterial enzymes, MycDAAO and GthDAAO, have different residues, Met, Phe, Ala, and Cys, in this position. In addition, these two features cannot be used to annotate the enzyme as a DAAO, since the same pair (the fingerprint sequence and the pair of conserved ArgPro residues) is present in all glycine oxidases. When only bacterial DAAO sequences are aligned (*Fig. 2*), the situation is more optimistic. As follows from *Fig. 2*, the region of the conserved ArgPro pair for all DAAOss in bacterial enzymes expands to the GxRPxR sequence, and a new conserved sequence YGHGGxG also appears. However, some glycine oxidases also have such sequences (not shown in *Fig. 2*).

One of the sequences found belongs to DAAO from the archaea *N. halalkaliphilus* AArcht4 (NhaDAAO). *Figure 2* clearly shows that the amino acid sequence of this enzyme is longer than that of the bacterial DAAOs. The alignment clearly shows three large inserts in the region of the FAD- and substrate-binding domains, as well as at the C-terminus. However, NhaDAAO contains the same set of conserved residues as bacterial DAAOs. In the second analysed enzyme from *N. halalkaliphilus* AArcht4, glycine oxidase NhaGOX, the positions of insertions and deletions coincided very well with similar positions in GOX from *P. aeruginosa* (*Fig. 1*) and other glycine oxidases (not shown).

The results of the comparison of the amino acid sequences allow us to make some assumptions about the type of substrate specificity. If the size of the substrates differs significantly, one can easily notice differences in the length of the regions that form the substrate-binding domain, already at the amino acid sequence alignment stage. For example, TvaDAAO, RemDAAO, and FsoDAAO have longer sequences in the region of residues 100–108 (*Fig. 1*, left side of the second row of alignment). This is because TvaDAAO and FsoDAAO are able to oxidize bulk cephalosporin C, while other DAAOs that have deletions in this part of the alignment do not oxidize cephalosporin C. For example, CboDAAO is specific to small amino acids, primarily D-Ala [17]. However, it should be noted that a homology search does not distinguish the classical DAAO with a wide spectrum of substrate specificity from the D-amino acid oxidase DASPO, which is specific only to D-Asp and D-Glu. For example, OpaDAAO1 (*Table 1*) is listed in the *O. parapolymorpha* DL-1 genome annotation as D-aspartate oxidase (DASPO), although our experimental data indicate that this is absolutely not the case. Our results showed that the enzyme has a wide range of substrate specificity and is identical to RtoDAAO and TvaDAAO in pH profiles of activity and stability. A similar situation is observed in the annotation of the *Pichia pastoris* genome. PpaDAAO1, annotated as a D-amino acid oxidase, is actually DASPO, while PpaDAAO2, annotated as a “hypothetical protein with low homology to D-amino acid oxidase,” is exactly DAAO [14].

Construction of model 3D structures and their comparative analysis with known DAAO structures

As already noted in the Introduction, the goal of searching for and cloning new genes is not just to obtain a recombinant enzyme but to create a biocatalyst with desired properties using the closest enzyme to the target as the initial one. In the case of DAAO, it is simply impossible to draw a conclusion about the properties (primarily about the substrate specificity and the optimal pH profile of activity) based on the alignment. In this regard, the use of additional methods is required. To solve this problem, we proposed an approach based on 3D structure modelling. In the first stage, model structures of new enzymes are compared with experimental and model structures of known D-amino acid oxidases and glycine oxidases.

Many examples have been published where enzymes with low homology have very close spatial structures. A good example is the supersecondary structure called the Rossman fold, which is universal for binding the adenine part of various cofactors and coenzymes, such as $NAD(P)^+$, FAD, ATP, SAM, etc. [31]. Using this approach to DAAO until recently was impossible due to the lack of a representative set of structures. Experimental structures were solved for only four enzymes: yeast RtoDAAO and RemDAAO and enzymes from pig kidney (pkDAAO) and humans (hDAAO). A model structure of TvaDAAO was constructed [32]. However, this enzyme is very similar to RemDAAO in both its primary (*Fig. 1*) and tertiary (*Table 2*) structures. In addition, the previously used modelling methods gave good results only at high sequence homology between the studied enzyme and the enzyme whose structure is used as a template for constructing a 3D model structure. High accuracy was achieved with a homology of at least 50–60%, which is not observed in the case of DAAO. The situation changed dramatically when a new algorithm for model structure construction, AlphaFold, was proposed in 2021 [33]. In 2022, the prediction accuracy was significantly improved [19]. The use of AlphaFold2 makes it possible to obtain reliable information on the structure of both new enzymes and already described DAAOs. Such model structures were constructed in our work. The structures of 18 DAAOs (including eight new ones) have been modelled. *Table 2* shows the results of a pairwise comparison of model and experimental structures of D-amino acid oxidases and glycine oxidases. In this case, the set of analysed DAAO structures was extended with two

Table 2. The standard deviation between the structures of D-amino acid oxidases and glycine oxidases*

Enzyme	**OpaDAAO1**	**OpaDAAO2**	**OpaDAAO3**	**OpaDAAO4**	**OpaDAAO5**	**OpaDASPO**	**TvaDAAO_RSA**	**OpaDAAO1_RSA**	**RtoDAAO_1C0P**	**RemDAAO_7CT4**	**hDAAO_2DU8**	**pkDAAO_1KIF**	**GOX_1NG4**	**IDA_OX_6PXS**	***NhaDAAO***	***NhyDAAO***	***NalGOX***	***NhaGOX***	***RraDAAO***	***PaeGOX***	***RbaDAAO***	***CthDAAO***	***RtaDAAO***	***GthDAAO***
OpaDAAO1	0.00	0.79	0.43	1.19	1.24	1.37	0.89	0.38	0.93	1.00	1.40	1.33	6.22	7.72	3.50	0.98	2.79	3.13	1.05	4.61	1.23	0.80	0.92	1.52
OpaDAAO2	0.79	0.00	0.71	0.81	0.74	1.06	0.66	0.82	0.84	0.79	1.46	1.34	5.51	10.29	2.22	0.91	7.72	4.31	0.80	5.66	0.87	0.86	0.79	0.96
OpaDAAO3	0.43	0.71	0.00	1.14	1.12	1.13	0.89	0.53	1.15	0.83	1.53	2.74	6.20	12.49	4.69	1.00	5.01	13.81	0.82	7.17	1.10	0.88	0.82	1.19
OpaDAAO4	1.19	0.81	1.14	0.00	0.96	1.30	1.01	1.21	1.17	1.08	2.00	2.33	3.10	21.04	2.21	1.10	3.94	2.79	0.99	3.07	1.31	1.36	1.27	1.15
OpaDAAO5	1.24	0.74	1.12	0.96	0.00	1.45	0.97	1.16	1.20	1.07	2.99	2.96	3.31	19.18	3.90	1.14	3.72	11.66	1.04	5.26	1.53	1.12	1.16	1.70
OpaDASPO	1.37	1.06	1.13	1.30	1.45	0.00	1.19	1.36	1.01	1.15	1.58	2.30	2.85	6.44	3.24	0.95	6.71	3.98	1.00	3.76	1.14	1.18	1.04	1.03
TvaDAAO_RSA	0.89	0.66	0.89	1.01	0.97	1.19	0.00	0.78	0.88	0.73	1.98	1.94	4.33	6.27	3.54	0.95	7.32	4.60	0.84	11.75	0.98	1.21	0.96	1.05
OpaDAAO1_RSA	0.38	0.82	0.53	1.21	1.16	1.36	0.78	0.00	0.98	0.92	1.45	1.35	5.95	9.63	3.46	1.04	2.94	4.25	1.04	4.87	1.23	0.88	0.93	1.49
RtoDAAO_1C0P	0.93	0.84	1.15	1.17	1.20	1.01	0.88	0.98	0.00	0.88	1.47	1.52	3.66	3.85	2.05	0.89	7.28	5.70	0.86	3.65	0.86	0.82	0.93	1.16
RemDAAO_7CT4	1.00	0.79	0.83	1.08	1.07	1.15	0.73	0.92	0.88	0.00	1.56	1.94	4.81	15.72	3.09	0.91	11.83	2.71	0.78	3.79	0.99	1.01	0.96	1.15
hDAAO_2DU8	1.40	1.46	1.53	2.00	2.99	1.58	1.98	1.45	1.47	1.56	0.00	0.41	4.50	4.52	1.96	1.09	8.68	9.48	1.16	8.13	1.10	1.38	1.12	1.12
pkDAAO_1KIF	1.33	1.34	2.74	2.33	2.80	2.30	1.94	1.35	1.52	1.94	0.41	0.00	3.78	4.22	2.83	1.05	4.76	4.67	1.08	5.36	1.07	1.31	1.05	1.07
GOX_1NG4	6.12	5.51	6.20	3.10	3.31	2.85	4.33	6.11	3.66	4.81	4.50	3.78	0.00	1.33	17.38	2.31	1.24	1.01	2.15	1.21	3.29	3.56	5.44	2.78
IDA_Ox_6PXS	7.72	10.29	12.49	21.04	19.18	6.44	6.27	9.63	3.85	14.20	5.03	4.22	1.33	0.00	9.26	4.28	1.13	0.91	2.67	1.45	6.90	2.69	5.01	3.35
NhaDAAO	3.50	2.25	4.69	2.21	3.90	3.24	3.54	3.46	2.05	3.09	1.96	2.83	17.38	9.26	0.00	1.19	9.48	5.26	0.88	10.28	2.59	1.67	2.01	3.00
NhyDAAO	0.98	0.91	1.00	1.10	1.14	0.95	0.95	1.04	0.89	0.91	1.09	1.05	2.31	4.28	1.19	0.00	1.95	3.09	0.56	2.19	0.50	0.71	0.46	0.70
NalGOX	2.79	7.72	5.01	3.94	3.72	6.25	7.32	2.94	7.28	12.83	8.68	4.76	1.24	1.13	16.19	1.95	0.00	1.21	3.99	1.54	1.51	2.43	3.79	5.80
NhaGOX	3.13	4.31	13.81	2.79	11.66	3.98	4.60	4.25	5.70	2.71	9.48	4.67	1.01	0.91	5.26	3.09	1.21	0.00	2.95	1.16	11.37	5.12	5.19	6.13
RraDAAO	1.05	0.80	0.82	0.99	1.04	1.00	0.84	1.04	0.86	0.78	1.16	1.08	2.15	2.67	0.88	0.56	3.99	2.95	0.00	2.33	0.52	0.59	0.41	0.61
PaeGOX	4.61	5.66	7.16	3.07	5.26	3.76	11.75	4.87	3.65	3.79	8.13	5.36	1.21	1.45	10.28	2.19	1.54	1.16	2.33	0.00	3.31	3.67	3.15	2.96
RbaDAAO	1.23	0.87	1.10	1.31	1.53	1.14	0.98	1.23	0.86	0.99	1.10	1.07	3.29	6.90	2.59	0.50	1.51	11.37	0.52	3.31	0.00	0.56	0.45	0.78
CthDAAO	0.80	0.86	0.88	1.36	1.12	1.18	1.21	0.88	0.82	1.01	1.38	1.31	3.56	2.59	1.67	0.71	2.43	5.12	0.59	3.67	0.56	0.00	0.63	0.94
RtaDAAO	0.92	0.79	0.82	1.27	1.16	1.04	0.96	0.93	0.93	0.96	1.12	1.05	5.44	5.01	2.01	0.46	3.79	5.19	0.41	3.15	0.45	0.63	0.00	0.64
GthDAAO	1.52	0.96	1.19	1.15	1.70	1.03	1.05	1.49	1.16	1.15	1.13	1.07	2.78	3.35	3.00	0.70	5.80	6.13	0.61	2.96	0.78	0.94	0.64	0.00

*New DAAO sequences analysed in this study are shown in bold italics.

experimental mammalian DAAO structures, from pig and human kidney, as well as with the structures of two glycine oxidases. In addition, our own preliminary data from the X-ray diffraction analysis of TvaDAAO and OpaDAAO1 were also used for comparison. Such an extended set allows more accurate comparison and increases the reliability of assigning new proteins to DAAO or GOX. For convenience, the comparison results shown in *Table 2* are highlighted in color. The green background shows the results comparing structures with RMSD up to 1 Å, light green with RMSD from 1 to 2 Å, light orange with RMSD from 2 to 6 Å, and orange with RMSD above 6 Å. Several important and interesting results of the analysis of the data in *Table 2* can be noted.

(1) The latest modification of the AlphaFold algorithm in the 2022 version [19] truly allows one to obtain model structures with very high accuracy. This is clearly seen when comparing the model and experimental structures of OpaDAAO1. The RMSD between these structures is only 0.38 Å. The RMSD between the model and experimental TvaDAAO structures is slightly larger, 0.56 Å (not shown in *Table 2*), but it should be taken into account that these enzymes have different oligomeric structures (OpaDAAO1 is a monomer, TvaDAAO is a dimer). The high accuracy of OpaDAAO1 structure prediction leads to the fact that a pairwise comparison of the model and experimental structures of OpaDAAO1 with the model and experimental structures of other enzymes gives almost identical RMSD values (*Table 2*, lines 1 and 8).

(2) There is a clear correlation between the function and overall structure of D-amino acid oxidases and glycine oxidases. The value of RMSD deviation between DAAO structures does not exceed 2 Å, while when comparing DAAO and GOX structures, the RMSD value is 3 Å or more (up to 15–18 Å). The results with NhaDAAO from archaea slightly fall out of the general picture – the deviation of the model structure from the structures of other DAAOs is 2.0–3.5 Å (in the case of OpaDAAO3, the deviation reaches even 4.69 Å). At the same time, the difference in the structure of NhaDAAO with glycine oxidases is much greater, from 9 to 17 Å. We should also note that this enzyme has a general structure close to that of human DAAO (RMSD is only 1.96 Å). Such results indicate that to correctly confirm that this enzyme is a DAAO, as broad a set of structures of known D-amino acid oxidases as possible should be used. Nevertheless, although the results of a general comparison of the structure of NhaDAAO with the structures of other DAAOs are generally slightly outside the 2 Å boundary value, the homology analysis and comparison of the general structure allowed us to classify this enzyme as a D-amino acid oxidase. The results of comparison of the structures of active centers fully confirm this conclusion.

Comparative Analysis of the Structures of DAAO Active Centers

In the next step, we compared the structures of the active centers of the new DAAOs with the known D-amino acid oxidases. The coincidence of the structure of the new protein with the structure of the active center of known enzymes clearly shows that the protein of interest belongs to certain enzyme families. The structure of the FAD-binding domain should be very similar in all DAAOs, but due to different specificities, the structure of substrate-binding domains should differ quite significantly both in volume and in the type of residues involved in the binding of a particular D-amino acid. Therefore, the coincidence of the structures of the substrate-binding domains of the active center allows one to unambiguously prove that the new enzyme belongs to the DAAO family and to draw a fairly reliable conclusion about the possible spectrum of substrate specificity. In this case, a comparison with the structures of DAAO active centers from the yeast *O. parapolymorpha* DL-1 is particularly useful since these enzymes differ greatly from each other both in the profile of substrate specificity and in the pH dependences of activity and stability. The effectiveness of such a comparison is clearly seen by the example of NhaDAAO from archaea. As noted above, this enzyme differs rather markedly from other DAAOs both in its amino acid sequence length and in its overall structure. However, the results of a comparison of the structures of the active sites indicate that NhaDAAO and OpaDAAO2 have almost identical active centers (*Fig. 3*). Alignment of the overall structures with the FAD cofactor reveals that in addition to the conserved Arg residue in the substrate-binding domain (see above), there are Tyr and Phe residues involved in substrate binding, and the locations of these residues in the active centers of NhaDAAO and OpaDAAO2 are almost identical. Moreover, the results of modelling the structure of the active center of OpaDAAO2 itself are in complete agreement with the experimental data, according to which the best substrates are D-amino acids with hydrophobic side groups – D-Phe (the highest activity), D-Tyr and D-Leu. Therefore, it is logical to assume that NhaDAAO should have the same spectrum of substrate specificity. High specificity to D-Leu and D-Phe was also predicted by comparison with the active center of OpaDAAO3 and the enzyme from *N. hydrolyticum* ACPA39 (NhyDAAO) (not shown). At present, the gene of this enzyme has been cloned

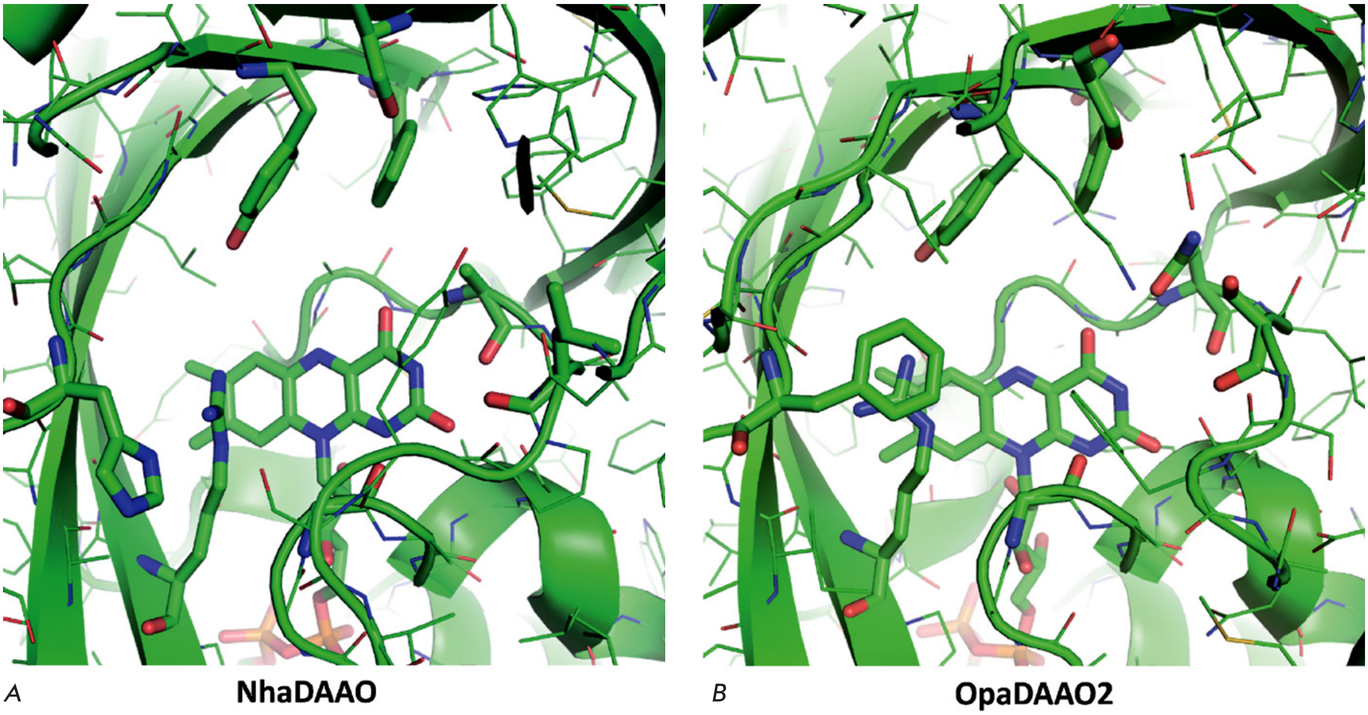


Fig. 3. The model structures of the active sites of *NhaDAAO* from the archaea *N. halalkaliphilus* AArcht4 (*A*) and OpaDAAO2 from the methylotrophic yeast *O. parapolymorpha* DL-1 (*B*)

in our laboratory, and its expression in *E. coli* cells is in progress. Preliminary experiments confirmed that D-Leu and D-Phe are the best substrates for NhyDAAO (a detailed description of the preparation and study of the properties of recombinant NhyDAAO will be presented in a separate publication).

A comparison of DAAO structures showed that the active centers of enzymes from *G. thermophila* (GthDAAO) and from *R. radiotolerans* DSM 5868 (RraoDAAO) are quite unique. In GthDAAO, the carboxyl groups of the side chain of residues Glu202 and Asp204 must participate in substrate binding (*Fig. 4A*). This suggests that this enzyme can be specific to D-Lys and D-Arg, but docking of various D-amino acids indicates that both carboxyl groups of residues Glu202 and Asp204 are located quite far away (more than 3 Å) from the substrate molecule. A more interesting situation is observed in the case of RraDAAO (*Fig. 4B*). The positively charged residue Arg226 and the negatively charged residue Glu228 can participate in the binding of the substrate side groups. Docking to the active center of various D-amino acids suggests that RraDAAO should be specific to positively charged D-Lys and potentially active with D-Glu. Cloning of the gene for this enzyme is of interest because D-Lys is a poor substrate for all DAAOs described.

CONCLUSION

The results of our experiments allow us to draw several conclusions.

(1) The introduction of the second stage – structural analysis – in the identification of genes for new D-amino acid oxidases, after a search in genomes by homology, is a highly effective and necessary procedure. At this stage, it is possible not only to unambiguously confirm that the new enzyme belongs to DAAO but also to predict the possible spectrum of its substrate specificity. The reliability of such prediction of high activity with D-Leu and D-Phe for new DAAO from the bacterium from *N. hydrolyticum* ACPA39 was confirmed experimentally.

(2) The amino acid sequences of D-amino acid oxidases from bacteria have low homology (no more than 30%). Analysis of the bacterial DAAO sequences revealed new characteristic conserved elements that can be used for identification of these enzymes during their search in bacterial genomes. The presence of new conserved regions was also shown in the DAAO sequence of *N. halalkaliphilus* AArcht4 (NhaDAAO) archaea.

(3) The D-amino acid oxidase gene was found in the archaean genome for the first time. Compared to bacterial DAAOs, the NhaDAAO enzyme from ar-

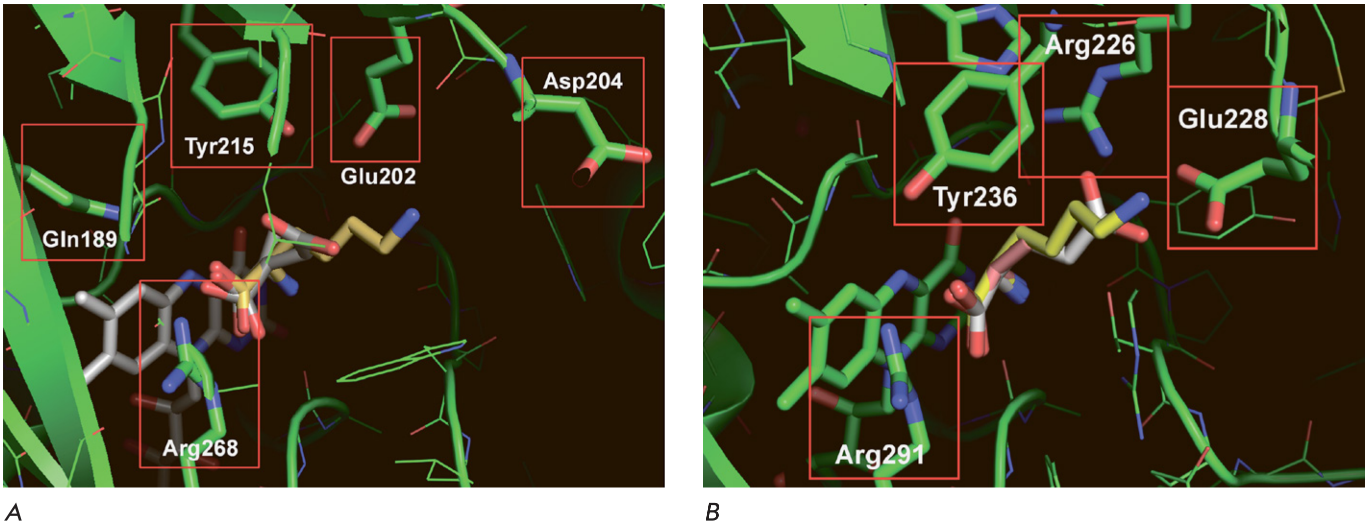


Fig. 4. Docking of D-Ala, D-Asp and D-Lys to the DAAO active site from *G. thermophila* (*A*) and D-Ala, D-Glu and D-Lys to the DAAO active site from *R. radiotolerans* DSM 5868 (*B*)

chaea has a longer amino acid sequence and less similar overall three-dimensional structure, but the results of structural analysis clearly showed that the active center of NhaDAAO is almost identical to the active center of OpaDAAO2 from the methylotrophic yeast *O. parapolymorpha* DL-1. Additionally, a glycine oxidase was identified in the genome of *N. halalkaliphilus* AArcht4, which is the closest in homology to GOX from the pathogen *P. aeruginosa*.

(4) D-amino acid oxidases play an important role in the functioning of microorganisms and mammals. That is why the search for human hDAAO inhibitors is one of the most active and topical areas of research on this enzyme [34]. Reliable identification of the D-amino acid oxidase gene (MycDAAO) in the genome of the tuberculosis causative pathogen allows us to consider this enzyme as a target for the development of a new type of drug against tuberculosis. Due to the rare occurrence of DAAO in bacteria and due to the significant differences of this enzyme from other DAAOs (primarily from hDAAO), efficient inhibitors that bind specifically to MycDAAO can be used as anti-tuberculosis drugs. ●

The authors declare no conflict of interest.

Searching for the genes of novel D-amino acid oxidases in the genomes of extremophilic bacteria and archaebacteria was carried out under Agreement No. 075-15-2021-1396 dated October 26, 2021, on allocating grant funds from the federal budget for implementing selected events under the Federal Research Program for Genetic Technologies Development for 2019–2027. Gene cloning, expression, isolation, characterization, and building model structures of enzymes from yeast O. parapolymorpha DL-1 were conducted under the grant from the Russian Foundation for Basic Research (RFBR grant No. 21-34-70040 mol_a_mos). Sequence analysis of enzymes from pathogenic microorganisms was carried out under State Assignment.

REFERENCES

1. Tishkov V.I., Khoronenkova S.V. // Biochemistry (Moscow). 2005. V. 70. № 1. P. 40–54. doi: 10.1007/s10541-005-0050-2
2. Pollegioni L., Piubelli L., Sacchi S., Pilone M.S., Molla G. // Cell Mol. Life Sci. 2007. V. 64. P. 1373–1394. doi: 10.1007/s00018-007-6558-4
3. Chumakov I., Blumenfeld M., Guerassimenko O., Cavarec L., Palicio M., Abderrahim H., Bougueleret L., Barry C., Tanaka H., La R.P., et al. // Proc. Natl. Acad. Sci. USA. 2002. V. 99. P. 13675–13680. doi: 10.1073/pnas.182412499
4. Cheng Y.J., Lin C.H., Lane H.Y. // Int. J. Mol. Sci. 2021. V. 22(20). P. 10917. doi: 10.3390/ijms222010917
5. Pernot P., Mothet J.P., Schuvailo O., Soldatkin A., Pollegioni L., Pilone M., Adeline M.T., Cespuglio R., Marinesco S. // Anal. Chem. 2008. V. 80. P. 1589–1597. doi: 10.1021/ac702230w
6. Khoronenkova S.V., Tishkov V.I. // Biochemistry (Moscow). 2008. V. 73. P. 1511–1518. doi: 10.1134/s0006297908130105
7. Pollegioni L., Molla G., Sacchi S., Rosini E., Verga R., Pilone M.S. // Appl. Microbiol. Biotechnol. 2008. V. 78.

P. 1–16. doi: 10.1007/s00253-007-1282-4
8. Pollegioni L., Molla G. // Trends Biotechnol. 2011. V. 29. P. 276–283. doi: 10.1016/j.tibtech.2011.01.010
9. Takahashi S., Abe K., Kera Y. // Bioengineered. 2015. V. 6. P. 237–241. doi: 10.1080/21655979.2015.1052917
10. Pollegioni L., Caldinelli L., Molla G., Sacchi S., Pilone M.S. // Biotechnol. Prog. 2004. V. 20. P. 467–473. doi: 10.1021/bp034206q
11. Atroshenko D.L., Shelomov M.D., Zarubina S.A., Negru N.Y., Golubev I.V., Savin S.S., Tishkov V.I. // Int. J. Mol. Sci. 2019. V. 20. P. 4412. doi: 10.3390/ijms20184412
12. Isogai T., Ono H., Ishitani Y., Kojo H., Ueda Y., Kohsaka M. // J. Biochem. 1990. V. 108. P. 1063–1069. doi: 10.1093/oxfordjournals.jbchem.a123306
13. Gonzalez F.J., Montes J., Martin F., Lopez M.C., Ferminan E., Catalan J., Galan M.A., Dominguez A. // Yeast. 1997. V. 13. P. 1399–1408. doi: 10.1002/(SICI)1097-0061(199712)13:15<1399
14. Pollegioni L., Molla G., Campaner S., Martegani E., Pilone M.S. // J. Biotechnol. 1997. V. 58. P. 115–123. doi: 10.1016/s0168-1656(97)00142-9
15. Klompmaker S.H., Kilic A., Baerends R.J., Veenhuis M., van der Klei I.J. // FEMS Yeast Res. 2010. V. 10. P. 708–716. doi: 10.1111/j.1567-1364.2010.00647.x
16. Yurimoto H., Hasegawa T., Sakai Y., Kato N. // Biosci. Biotechnol. Biochem. 2001. V. 65. P. 627–633. doi: 10.1271/bbb.65.627
17. Shimekake Y., Furuichi T., Abe K., Kera Y., Takahashi S. // Sci. Rep. 2019. V. 9. P. 11948. doi: 10.1038/s41598-019-48480-y
18. Atroshenko D., Shelomov M., Zhgun A., Avdanina D., Eldarov M., Pometun A., Chubar T., Savin S., Tishkov V. // FEBS Open Bio. 2018. V. 8(S1). P. 190. doi: 10.1002/2211-5463.12453
19. Jumper J., Hassabis D. // Nat. Methods. 2022. V. 19. P. 11–12. doi: 10.1038/s41592-021-01362-6
20. Mirdita M., Schutze K., Moriwaki Y., Heo L., Ovchinnikov S., Steinegger M. // Nat. Methods. 2022. V. 19. P. 679–682. doi: 10.1038/s41592-022-01488-1
21. Emsley P., Lohkamp B., Scott W.G., Cowtan K. // Acta Crystallogr. D. Biol. Crystallogr. 2010. V. 66. P. 486–501. doi: 10.1107/S0907444910007493
22. Morris G.M., Huey R., Lindstrom W., Sanner M.F., Belew R.K., Goodsell D.S., Olson A.J. // J. Comput. Chem. 2009. V. 30. P. 2785–2791. doi: 10.1002/jcc.21256
23. Santos-Martins D., Solis-Vasquez L., Tillack A.F. // J. Chem. Theory. Comput. 2021. V. 17. P. 1060–1073. doi: 10.1021/acs.jctc.0c01006
24. Sorokin D.Y., Elcheninov A.G., Khijniak T.V., Zaharycheva A.P., Boueva O.V., Ariskina E.V., Bunk B., Sproer C., Evtushenko L.I., Kublanov I.V., Hahnke R.L. // Appl. Microbiol. 2022. V. 45. P. 126307. doi: 0.1128/AEM.02193-14
25. Sorokin D.Y., Khijniak T.V., Zakharycheva A.P., Elcheninov A.G., Hahnke R.L., Boueva O.V., Ariskina E.V., Bunk B., Kublanov I.V., Evtushenko L.I. // Int. J. Syst. Evol. Microbiol. 2021. V. 71. P. 04804. doi: 10.1099/ijsem.0.004804
26. Sorokin D.Y., Elcheninov A.G., Toshchakov S.V., Bale N.J., Sinninghe Damste J.S., Khijniak T.V., Kublanov I.V. // Appl. Microbiol. 2019. V. 42. P. 309–318. doi: 10.1016/j.syapm.2019.01.001
27. Alekseeva A.A., Savin S.S., Tishkov V.I. // Acta Naturae. 2011. V. 3. P. 38–54. PMID: 22649703
28. Tishkov V.I., Popov V.O. // Biochemistry (Moscow). 2004. V. 69. P. 1252–1267. doi: 10.1007/s10541-005-0071-x.
29. Tishkov V.I., Popov V.O. // Biomol. Eng. 2006. V. 23. P. 89–110. doi: 10.1016/j.bioeng.2006.02.003
30. Baker P.J., Britton K.L., Rice D.W., Rob A., Stillman T.J. // J. Mol. Biol. 1992. V. 228. P. 662–671. doi: 10.1016/0022-2836(92)90848-e
31. Rao S.T. Rossmann M.G. // J. Mol. Biol. 1973. V. 76. P. 241–256. doi: 10.1016/0022-2836(73)90388-4
32. Tishkov V.I., Khoronenkova S.V., Cherskova, N.V., Savin S.S., Uporov I.V. // Moscow Univ. Chem. Bull. 2010. V. 65. № 3. P. 121–126. doi: 10.3103/S0027131410030028
33. Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Zidek A., Potapenko A., et al. // Nature. 2021. V. 596. P. 583–589. doi: 10.1038/s41586-021-03819-2
34. Pollegioni L., Sacchi S., Murtas G. // Front. Mol. Biosci. 2018. V. 5. P. 107. doi: 10.3389/fmolb.2018.00107

Rhodobacter capsulatus PG Lipopolysaccharide Blocks the Effects of a Lipoteichoic Acid, a Toll-Like Receptor 2 Agonist

S. V. Zubova[1*], N. I. Kosyakova[2], S. V. Grachev[1,3], I. R. Prokhorenko[1]
[1]Institute of Basic Biological Problems of RAS FRC PSCBR RAS, Pushchino, 142290 Russia
[2]Clinical Hospital at the Pushchino Research Center, Pushchino, 142290 Russia
[3]First Moscow State Medical University named I.M. Sechenov of Russia Health Ministry (Sechenov University), Moscow, 119991 Russia
*E-mail: zusvet@rambler.ru

Received June 07, 2022; in final form, November 02, 2022
DOI: 10.32607/actanaturae.11747

ABSTRACT Lipopolysaccharides (LPS) and lipoteichoic acids (LTA) are the major inducers of the inflammatory response of blood cells caused by Gram-negative and some Gram-positive bacteria. CD14 is a common receptor for LPS and LTA that transfers the ligands to TLR4 and TLR2, respectively. In this work, we have demonstrated that the non-toxic LPS from *Rhodobacter capsulatus* PG blocks the synthesis of pro-inflammatory cytokines during the activation of blood cells by *Streptococcus pyogenes* LTA through binding to the CD14 receptor, resulting in the signal transduction to TLR2/TLR6 being blocked. The LPS from *Rhodobacter capsulatus* PG can be considered a prototype for developing preparations to protect blood cells against the LTA of gram-positive bacteria.

KEYWORDS lipopolysaccharide, *Rhodobacter capsulatus*, lipoteichoic acid, TLR, CD14, cytokines.

ABBREVIATIONS CD – cluster of differentiation; ERK – extracellular signal-related kinase; IL – interleukin; JNK – c-Jun N-terminal kinase; LBP – LPS-binding protein; LPS – lipopolysaccharide; LTA – lipoteichoic acid; MAPK – mitogen-activated protein kinase; MD-2 – myeloid differentiation protein 2; NF-ϰB – nuclear factor kappa B; PAMP – pathogen-associated molecular patterns; PI3K – phosphatidylinositol 3 kinase; PKC – protein kinase C; TLR – Toll-like receptor; TNF-α – tumor necrosis factor α.

INTRODUCTION

Studying the mechanisms of inflammation induced by ligands of differing nature is one of the priorities in modern biomedicine. This work considers the possibility of using lipopolysaccharide (LPS) from *Rhodobacter capsulatus* PG, a non-toxic endotoxin antagonist, to study the mechanisms underwriting the functional responses of innate immunity cells to pathogen-associated molecular patterns (PAMP) of differing nature. LPS and lipoteichoic acids (LTA), the central elements of the cell wall of Gram-negative and Gram-positive bacteria, exhibit immunostimulatory activity. LPS are glycolipids with three structural domains: lipid A, core oligosaccharide, and O-antigen, and they are localized in the outer membrane of Gram-negative bacteria. LTA are amphiphilic di- and triacylated lipopeptides anchored on the outer side of the cytoplasmic membrane of Gram-positive bacteria. In some aspects, LTA can be considered the equivalent of LPS, which is responsible for the development of the septic shock induced by Gram-positive bacteria [1]. TLR4 and TLR2 when expressed on the surface of blood cells can recognize these biologically active molecules. TLR4 has been identified as a specific receptor for LPS, inducing the release of pro-inflammatory cytokines by monocytes and macrophages stimulated by endotoxins [2]. TLR2 recognizes the di- or triacylated LTA of Gram-positive bacteria by triggering the immune response [3, 4]. The LTA from *Streptococcus pyogenes*, *Staphylococcus aureus,* and *Streptococcus pneumonia* bind directly to TLR2 [5–7]. The blood LBP protein, which binds to LPS and transfers it as a monomer to the membrane-bound receptor CD14, then to MD-2 and TLR4, is involved in the delivery of LPS to the receptor [8]. LBP and CD14 are also involved in LTA delivery to TLR2 [4]. CD14 constitutes part of the multi-ligand receptor complex, mediating a variety of cellular responses related to signal transduction from TLR2 and TLR4 [9]. CD14 enhances the TLR2 activation by facilitating lipopeptide binding and TLR2 heterodimerization with TLR1 or TLR6. The activation of the TLR2/TLR6 complex by diacylated lipopeptides, particularly LTA, involves the CD36 receptor [10]. For TLR4 to function as an LPS receptor, the myeloid differentiation factor MD-2 is required [11]. MD-2 is physically associated with TLR2 but weaker than it is with TLR4 [12]. This accessory molecule has been shown to enhance the TLR2-mediated responses to LTA [13]. Unlike TLR4, which transmits signals as a homodimer $(TLR4)_2$ when responding to LPS, TLR2 forms a heterodimer with TLR6 or TLR1 when recognized by LTA [14, 15]. The cell wall bacterial components LTA and LPS trigger the intracellular signaling cascade through TLR2 and TLR4 via a similar signaling pathway, that activates the transcription factors NF-ϰB, PKC, PI3K, ERK, JNK, and p38 MAPK and synthesizes the pro-inflammatory cytokines TNF-α, IL-1β, IL-6 and chemokine IL-8 [16]. LPS from a wide range of non-enterobacterial bacteria activate the myeloid cell line via TLR2 [17, 18]. The features of the lipid A of these LPS include a presence of phosphorylated diglucosamine, the length of hydrocarbon chains of fatty acid residues different from the chain length of enterobacterial LPS, or branched acyl chains [19]. The non-toxic LPS of the Gram-negative phototrophic bacterium *Rhodobacter capsulatus* PG functions as an endotoxin antagonist [20, 21]. This LPS can block blood cell activation, resulting in a wide range of pro-inflammatory cytokines being released caused by endotoxins [22]. E5531, a synthetic analog of lipid A from *R. capsulatus*, suppresses TNF-α production by human blood monocytes activated by *E. coli* LPS 0111:B4 or *Staphylococcis faecalis* LTA, exhibiting almost no activity of its own [23].

The structure of the non-toxic lipid A of the LPS from *Rhodobacter capsulatus* includes diphosphorylethanolamine at C-1, phosphorylethanolamine at C-4', and an unsaturated fatty acid (12:1) in the disaccharide backbone [24]. These structural features of lipid A allowed us to hypothesize that *Rhodobacter capsulatus* PG LPS, similar to E5531, could compete with *S. pyogenes* LTA for TLR2 by blocking the activation of pro-inflammatory cytokine synthesis by blood cells.

EXPERIMENTAL

The research was performed on the whole blood of healthy volunteers of both sexes, with ages ranging from 25 to 30 years. All subjects gave written consent to participate in the study. The study protocol complies with the Declaration of Helsinki of the World Medical Association (2013) and was approved by the Local Ethics Committee of the Hospital of the Pushchino Scientific Center (No. 2 of April 10, 2014). Peripheral blood was collected under clinical conditions using vacutainers (Becton Dickinson and Company, United Kingdom) treated with sodium heparin (17 units/ml).

Activation of blood cells by LPS and LTA

We studied the effect of LPS and LTA on cytokine and chemokine synthesis by diluting blood in RPMI 1640 medium at a ratio of 1 : 10 and incubating with *E. coli* LPS 055:B5 (100 ng/ml), *S. enterica* serotype Typhimurium LPS (100 ng/ml), *S. pyogenes* LTA (1000 ng/mL) (Sigma-Aldrich, USA), or *Rhodobacter capsulatus* PG LPS (1000 ng/mL) in various combinations for 6 and 24 h at 37°C in 5% CO_2. The *Rhodobacter capsulatus* PG LPS was obtained according to the method described previously [25]. We determined the antagonistic effect of *Rhodobacter capsulatus* PG LPS against *E. coli* LPS, *S. enterica* LPS, or *S. pyogenes* LTA in various combinations by preincubating blood with *Rhodobacter capsulatus* PG LPS for 30 min, followed by the addition of LPS or LTA. To determine the role of the CD14 receptor in cell activation, we preincubated the blood with antibodies (Ab) to CD14 (2 μg/ml) (Purified Anti-human CD14 Clone M5E2, BioLegend, USA) for 30 min at 4°C and then added LPS or LTA. The samples were incubated for 6 and 24 h at 37°C in 5% CO_2. Once incubated, the blood cells were precipitated by centrifugation (300 g, 10 min). The supernatants were collected and stored at −20°C until the cytokine and chemokine contents were determined.

Cytokine and chemokine content

The content of cytokines and chemokines was determined using TNF-α, IL-6, IL-1β, and IL-8 ELISA kits (Vector-BEST, Russia) according to the manufacturer's protocol. The optical density of the samples was determined using a STAT FAX 3200 ELISA analyzer (Awareness Technology Inc., USA) at a wavelength of 450 nm.

Statistical analysis

Statistical processing and graphical representation of the results were performed using nonparametric statistics in Origin Pro 7.5 and Microsoft Office Excel 2010 (AtteStat plugin). The results were presented as values with upper and lower quartiles (IQR). The statistical significance of the differences between median values was determined by the Mann-Whitney test ($p < 0.05$).

RESULTS

E. coli LPS or *S. enterica* LPS stimulated significant, similarly high, production of the pro-inflammatory cytokines TNF-α (*Fig. 1*), IL-6 (*Fig. 2*), and IL-1β (*Fig. 3*), as well as the inflammatory chemokine IL-8 (*Fig. 4*), whose production significantly exceeded control values. LTA activation also resulted in the production of high levels of the cytokines and chemokines analyzed. The level of synthesis of the later cytokine IL-1β and chemokine IL-8 in response to *S. pyogenes* LTA exceeded the levels when activated by *E. coli* LPS or *S. enterica* LPS (*Fig. 3, 4*).

Non-toxic *Rhodobacter capsulatus* PG LPS at a concentration 10-fold higher than that of the *E. coli* and *S. enterica* endotoxins and at equal concentration with *S. pyogenes* LTA did not stimulate the cells to produce TNF-α, IL-6, and IL-1β (*Fig. 1–3*). The amount of chemokine IL-8 in the blood in response to *Rhodobacter capsulatus* PG LPS slightly increased compared to the control but was significantly lower than that during the activation of blood cells by endotoxins or *S. pyogenes* LTA (*Fig. 4*). The study of the ability of *Rhodobacter capsulatus* PG LPS to protect blood cells from the action of the *E. coli* and *S. enterica* endotoxins revealed that the *Rhodobacter capsulatus* PG LPS suppressed the synthesis of the TNF-α, IL-6, and IL-1β cytokines in the blood, with the blocking response to *S. enterica* LPS being stronger than that to *E. coli* LPS (*Fig. 1–3*).

No protective effect of *Rhodobacter capsulatus* PG LPS against the endotoxins was observed according to the IL-8 chemokine synthesis (*Fig. 4*). IL-8 is an important mediator of the host response to inflammation and infection [26]. It is assumed that the cell response to an exposure to bacterial agents and IL-8 synthesis is induced earlier than the IL-6 synthesis [27].

Upon the activation of the cells with *S. pyogenes* LTA, pre-incubation of blood with *Rhodobacter capsulatus* PG LPS resulted in a significant decrease in the synthesis of the pro-inflammatory cytokines TNF-α, IL-6 and IL-1β and chemokine IL-8 (*Fig. 1–4*). The data obtained suggest that the LPS from *Rhodobacter capsulatus* PG exhibit antagonistic activity not only against endotoxins, but also against the *S. pyogenes* LTA.

In the control samples, Ab to CD14 did not affect the activation of the TNF-α synthesis in blood cells (*Fig. 5*). Pre-incubation of blood with Ab to CD14, followed by the activation of *E. coli* LPS, *S. enterica* LPS, or *S. pyogenes* LTA cells more markedly reduced the TNF-α synthesis induced by LTA than by endotoxins.

DISCUSSION

Toll-like receptors (TLRs) activate the cells of the innate immune system by recognizing various microorganisms through pathogen-associated molecular patterns (PAMPs), particularly LPS of Gram-negative bacteria and LTA of Gram-positive bacteria. TLR4 receptors recognize LPS, the central inducers of the inflammatory responses induced by Gram-negative bacteria, and TLR2 recognizes LTA, the inducers of the inflammatory response triggered by Gram-positive bacteria [3]. Both receptors are capable of signaling by forming a homodimer $(TLR4)_2$ or a TLR2/TLR6 heterodimer, respectively. Variations in the number of acyl chains in endotoxin lipid A can attenuate signaling through TLR4 and alter the host's immune response to the pathogen [28]. TLR4/MD-2 recognizes hexaacylated *E. coli* lipid A as an agonist. The structural changes in the lipid A of other Gram-negative bacteria reduce their activity in the receptor complex, compared to hexaacylated lipid A. When examining the ability of E5531, a pentaacylated synthetic analog of lipid A of *Rhodobacter capsulatus*, to inhibit the binding of *E. coli* LPS to human monocytes, was calculated the affinity of E5531 to the cells to be 24 times lower than that of *E. coli* LPS [23]. We used *Rhodobacter capsulatus* PG LPS in concentrations 10-fold higher than those of endotoxins to block the effects of *E. coli* LPS or *S. enterica* LPS. The LPS of *Rhodobacter capsulatus* PG was found to block the synthesis of the pro-inflammatory cytokines TNF-α, IL-6, and IL-1β in the cells activated by *S. enterica* LPS stronger than *E. coli* LPS. The antagonistic activity of the LPS of *Rhodobacter capsulatus* PG against the *S. pyogenes* LTA was significantly stronger when equal weight concentrations of *Rhodobacter capsulatus* PG LPS and *S. pyogenes* LTA were used. The ability of *Rhodobacter capsulatus* PG LPS to protect the cells from activation cytokine synthesis by agonists was reduced in the series of *S. pyogenes* LTA > *S. enterica* LPS > *E. coli* LPS (*Fig. 1–3*). The CD14 receptor, involved not only in ligand recognition by the TLR4 and TLR2 receptors, but also in the activation of cytokine synthesis by the cells, plays a critical role in both LPS and LTA signal transduction [6, 29]. The CD14 receptors expressed on the cell surface bind with high affinity to the molecular ligands associated with various pathogens. Subsequently, CD14 transmits LPS to the TLR4/MD-2 signaling complex [30]; and LTA, to the TLR2/TLR6 complex [4]. CD14 and CD36 act as TLR2 coreceptors in the monocyte response to LTA. Blocking

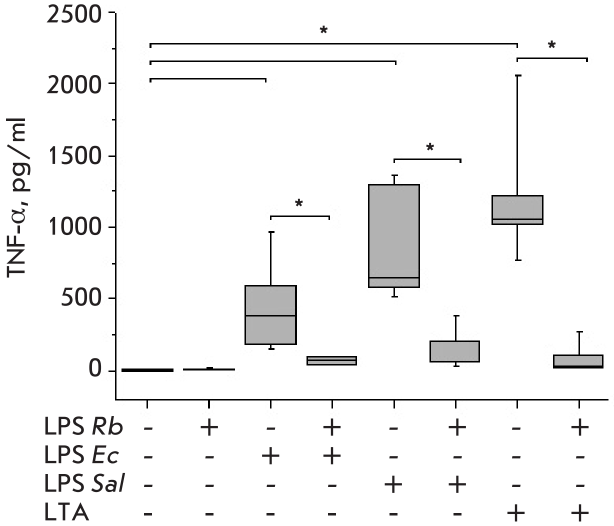


Fig. 1. Effect of *R. capsulatus* PG LPS on TNF-α secretion upon activation of blood cells by *E. coli* LPS, *S. enterica* LPS, or *S. piogenes* LTA, $n = 7$. $^*p < 0.05$

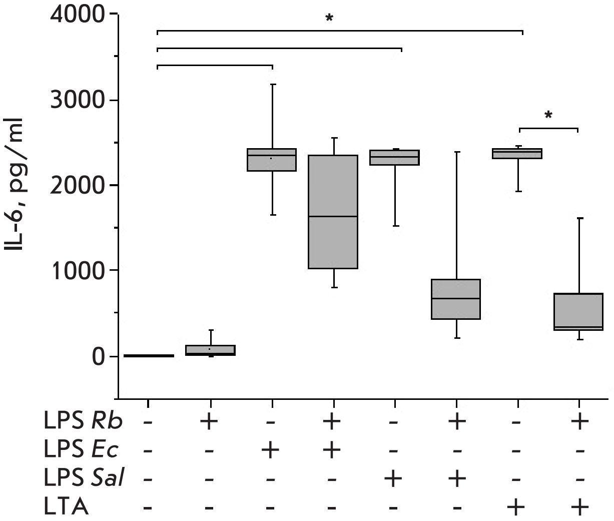


Fig. 2. Effect of *R. capsulatus* PG LPS on IL-6 secretion upon activation of blood cells by *E. coli* LPS, *S. enterica* LPS, or *S. piogenes* LTA, $n = 7$. $^*p < 0.05$

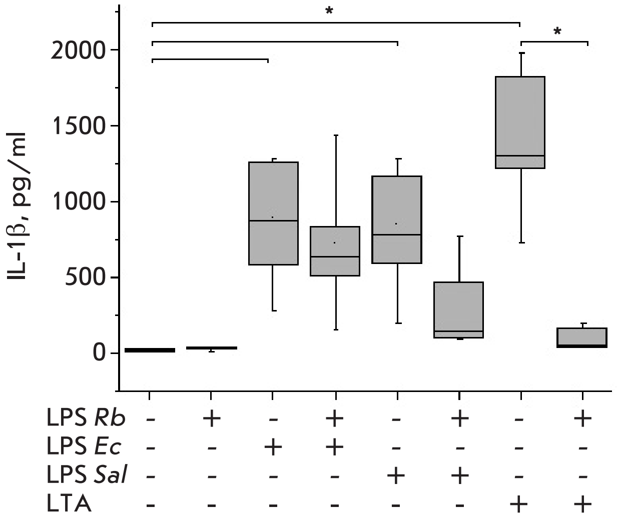


Fig. 3. Effect of *R. capsulatus* PG LPS on IL-1β secretion upon activation of blood cells by *E. coli* LPS, *S. enterica* LPS, or *S. piogenes* LTA, $n = 7$. $^*p < 0.05$

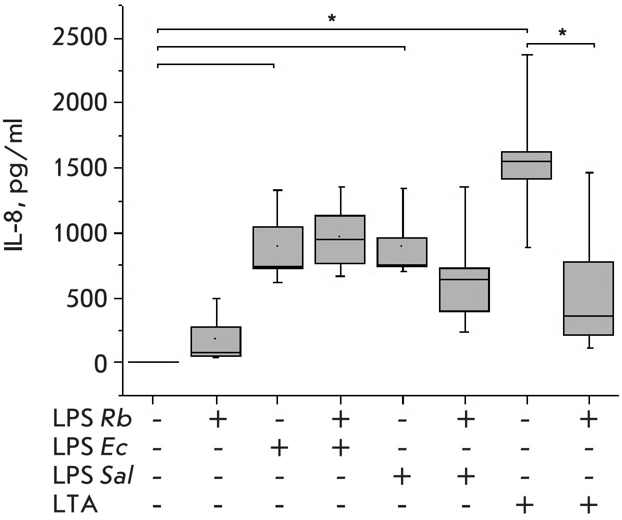


Fig. 4. Effect of *R. capsulatus* PG LPS on IL-8 secretion upon activation of blood cells by *E. coli* LPS, *S. enterica* LPS, or *S. piogenes* LTA, $n = 7$. $^*p < 0.05$

these receptors with antibodies inhibits the LTA-induced release of TNF-α by monocytes, indicating the involvement of these receptors in LTA binding to the plasma membrane and NF-ϰB activation [31]. On human monocytes, *Streptococcus sanguis* LTA has been shown to compete with *Salmonella abortusequi* LPS for binding to CD14. However, the LPS binding to CD14 has been found to be completely inhibited if the LTA concentration is 100-fold higher than the LPS concentration [32].

To validate this assumption and understand the mechanism of suppression of cell activation by

Rhodobacter capsulatus PG LPS, we blocked blood cell CD14 receptors using mAbs prior to activation by the LPS or LTA agonist. The low percentage of activation reduction observed (compared to the data in [23]) upon blocking of the CD14 receptors is obviously related to the specificity of the antibodies we used (Clone M5E2). The pre-incubation of blood with Ab CD14 before the activation of the cells by *E. coli* LPS, *S. enterica* LPS, or *S. pyogenes* LTA more markedly reduced the TNF-α synthesis induced by LTA than by the endotoxins. The results obtained demonstrate that CD14 is involved in the activation and signal transduction to cytokine synthesis from LPS and LTA, with this involvement decreasing in the series of *S. pyogenes* LTA > *S. enterica* LPS > *E. coli* LPS (*Fig. 5*), similar to the decreasing efficiency of *Rhodobacter capsulatus* PG LPS protection from cell activation by the agonists used (*Fig. 1-3*).

Two possible mechanisms for blocking cell activation by *Rhodobacter capsulatus* PG LPS can be suppose here. They are related to the different affinities of the studied ligands for the CD14 receptors: blocking at the level of interaction with the CD14 receptor or at the level of activation of the TLR4/MD-2 or TLR2/TLR6 receptor complex.

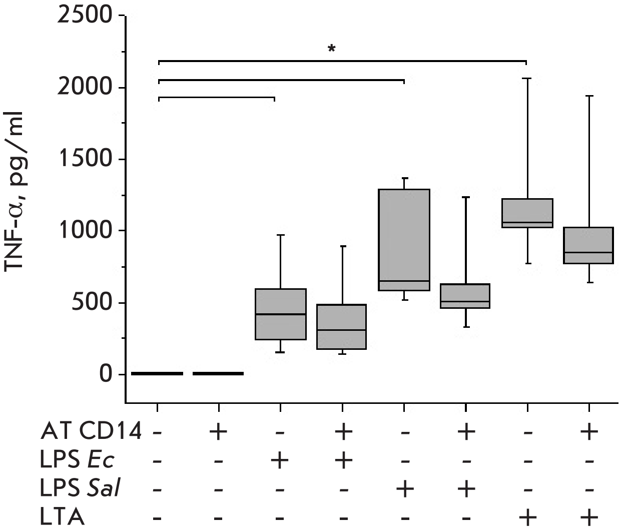


Fig. 5. Effect of Ab CD14 on TNF-α secretion upon activation of blood cells by *E. coli* LPS, *S. enterica* LPS, *S. pyogenes* LTA, $n = 7$. $^{*}p < 0.05$

CONCLUSION

Our results have revealed that the non-toxic LPS of *Rhodobacter capsulatus* PG blocks the synthesis of pro-inflammatory cytokines upon blood cell activation by *S. pyogenes* LTA through binding to the CD14 receptor, resulting in a suppression of signal transduction to TLR2/TLR6. To conclude, we believe that the LPS of *Rhodobacter capsulatus* PG can be considered a prototype for developing preparations to protect blood cells from the action of LTA of Gram-positive bacteria. ●

The authors would like to express their gratitude to the Hospital of the Pushchino Scientific Center of the Russian Academy of Sciences for cooperation.

This work was performed within the framework of State Assignment 122041200039-0 established by the Ministry of Education and Science of the Russian Federation at the Institute of Fundamental Problems of Biology, Russian Academy of Sciences.

The authors declare no conflict of interest.

REFERENCES

1. Ginsburg L. // Lancet. Infect. Dis. 2002. V. 2. № 3. P. 171–179.
2. Zhang G., Meredith T.C., Kahne D. // Curr. Opin. Microbiol. 2013. V. 16. № 6. P. 779–785.
3. Schwandner R., Dziarski R., Wesche H., Rothe M., Kirschning C.J. // J. Biol. Chem. 1999. V. 274. № 25. P. 17406–17409.
4. Schroder N.W.J., Morath S., Alexander C., Hamann L., Hartung T., Zahringer U., Gobel U.B., Weber J.R., Schumann R.R. // Biol. Chem. 2003. V. 278. № 18. P. 15587–15594.
5. Im J., Choi H.S., Kim S.K., Woo S.S., Ryu Y.H., Kang S.S., Yun C.H., Han S.H. // Cancer Lett. 2009. V. 274. № 1. P. 109–117.
6. Kang J.Y., Nan X., Jin M.S., Youn S.J., Ryu Y.H., Mah S., Han S.H., Lee H., Paik S.G., Lee J.O. // Immunity. 2009. V. 31. № 6. P. 873–884.
7. Fieber C., Janos M., Koestler T., Gratz N., Li X-D., Castiglia V., Aberle M., Sauert M., Wegner M., Alexopoulou L., et al. // PLoS One. 2015. V. 10. № 3. P. e0119727.
8. Ryu J.-K., Kim S.J., Rah S.-H., Kang J.I., Jung H.E., Lee D., Lee H.K., Lee J.-O., Park B.S., Yoon T.-Y., Kim H.M. // Immunity. 2017. V. 46. V. 1. P. 1–13.
9. Schmitz G., Orso E. // Curr. Opin. Lipidol. 2002. V. 13. № 5. P. 513–521.
10. Triantafilou M., Gamper F.G., Haston R.M., Mouratis M.A., Morath S., Hartung T., Triantafilou K. // Biol. Chem. 2006. V. 281. № 41. P. 31002–31011.
11. Shimazu R., Akashi S., Ogata H., Nagai Y., Fukudome K., Miyake K., Kimoto M. // J. Exp. Med. 1999. V. 189. № 11. P. 1777–1782.

12. Dziarski R., Wang Q., Miyake K., Kirsching C.J., Gupta D.J. // J. Immunol. 2001. V. 166. № 3. P. 1938–1944.
13. Dziarski R., Gupta D.J. // Endotox. Res. 2000. V. 6. № 5. P. 401–405.
14. Ozinsky A., Underhill D.M., Fontenot J.D., Hajjar A.M., Smith K.D., Wilson C.B., Schroeder L., Aderem A. // Proc. Natl. Acad. Sci. USA. 2000. V. 97. № 25. P. 13766–13771.
15. Henneke P., Morath S., Uematsu S., Weichert S., Pfitzenmaier M., Takeuchi O., Müller A., Poyart C., Akira S., Berner R., et al. // Immunol. 2005. V. 174. № 10. P. 6449–6455.
16. Su S.-H., Hua K.-F., Lee H., Chao L.K., Tan S.-K., Lee H., Yang S.-F., Hsu H.-Y. // Clin. Chim. Acta. 2006. V. 374. № 1–2. P. 106–115.
17. Yokota S., Ohnishi T., Muroi M., Tanamoto K., Fujii N., Amano K. // FEMS Immunol. Med. Microbiol. 2007. V. 51. № 1. P. 140–148.
18. Girard R., Pedron T., Uematsu S., Balloy V., Chignard M., Akira S., Chaby R. // J. Cell Sci. 2002. V. 116. Pt 2. P. 293–302.
19. Erridge C., Pridmore A., Eley A., Stewart J., Poxton I.R. // J. Med. Microbiol. 2004. V. 53. Pt 8. P. 735–740.
20. Katzke N., Bergmann R., Jaeger K.-E., Drepper T. // Methods Mol. Biol. 2012. V. 824. P. 251–269.
21. Prokhorenko I.R., Grachev S.V., Zubova S.V. Patent for invention RU № 2392309 of 20.06.2010.
22. Kabanov D.S., Serov D.A., Zubova S.V., Grachev S.V., Prokhorenko I.R. // Biochemistry (Moscow). 2016. V. 81. № 3. P. 275–283.
23. Kawata T., Bristol J.R., Rossignol D.P., Rose J.R., Kobayashi S., Yokohama H., Ishibashi A., Christ W.J., Katayama K., Yamatsu I., Kishi Y. // Br. J. Pharmacology. 1999. V. 127. № 4. P. 853–862.
24. Krauss J.H., Seydel U., Weckesser J., Mayer H. // Eur. J. Biochem. 1989. V. 180. № 3. P. 519–526.
25. Makhneva Z.K., Vishnevetskaya T.A., Prokhorenko I.R. // Pric. Biochim. Microbe. 1996. V. 32. № 4. P. 444–447.
26. Baggiolini M., Walz A., Kunkel S.L. // J. Clin. Invest. 1989. V. 84. № 4. P. 1045–1049.
27. Hirao Y., Kanda T., Aso Y., Mitsuhashi M., Kobayashi I. // Lab. Med. 2000. V. 31. № 1. P. 39–44.
28. Kawai T., Akira S. // Nat. Immunol. 2010. V. 11. № 5. P. 373–384.
29. Zanoni I., Ostuni R., Marek L.R., Baressi S., Barbalat R., Barton G.M., Granucci F., Kagan J.C. // Cell. 2011. V. 147. № 4. P. 868–880.
30. Wright S.D., Ramos R.A., Tobias P.S., Ulevitch R.J., Mathison J.C. // Science. 1990. V. 249. № 4975. P. 1431–1433.
31. Nilsen N.J., Deininge S., Nonstad U., Skjeldal F., Husebye H., Rodionov D., von Aulock S., Hartung T., Lien E., Bakke O., Espevik T.J. // Leukoc. Biol. 2008. V. 84. № 1. P. 280–291.
32. Sugawara S., Arakaki R., Rikiishi H., Takada H. // Infect. Immunol. 1999. V. 67. № 4. P. 1623–1632.

Comprehensive Analysis of Stromal and Serum Markers in Gastric Cancer

O. V. Kovaleva[1*], P. A. Podlesnaya[1], V. L. Chang[2], N. A. Ognerubov[2], A. N. Gratchev[1], N. A. Kozlov[1], I. S. Stilidi[1], N. E. Kushlinskii[1]

[1]N.N. Blokhin National Medical Research Center of Oncology of the Ministry of Health of Russian Federation, Moscow, 115552 Russia

[2]Medical Institute of G.P. Derzhavin Tambov State University, Tambov, 392000 Russia

[*]E-mail: ovkovaleva@gmail.com

Received June 14, 2022; in final form, October 20, 2022

DOI: 10.32607/actanaturae.11753

ABSTRACT A comprehensive analysis of the cell phenotype of the inflammatory infiltrate of the tumor stroma represents a promising area of molecular oncology. The study of not only soluble forms of various immunoregulatory molecules, but also their membrane-bound forms is also considered highly relevant. We performed a comprehensive analysis of tissue and circulating forms of the PD-1 and PD-L1 proteins, as well as macrophage and B-cell markers in the tumor stroma of gastric cancer, to assess their clinical and prognostic significance. The tumor and blood plasma samples from 63 gastric cancer patients were studied using ELISA and immunohistochemistry. Malignant gastric tumors were shown to be strongly infiltrated by B-cells, and their number was comparable to that of macrophages. For PU.1 expression, an association with tumor size was observed; i.e., larger tumors were characterized by fewer PU.1+ infiltrating cells ($p = 0.005$). No clinical significance was found for CD20 and CD163, but their numbers were higher at earlier stages of the disease and in the absence of metastases. It was also demonstrated that the PD-L1 content in tumor cells was not associated with the clinical and morphological characteristics of GC. At the same time, PD-L1 expression in tumor stromal cells was associated with the presence of distant metastases. The analysis of the prognostic significance of all the markers studied demonstrated that CD163 was statistically significantly associated with a poor prognosis for the disease ($p = 0.019$). In addition, PD-L1 expression in tumor cells tended to indicate a favorable prognosis ($p = 0.122$). The results obtained in this work indicate that the study of soluble and tissue markers of tumor stroma is promising in prognosticating the course of GC. The search for combinations of markers seems to be highly promising, with their comprehensive analysis capable of helping personalize advanced antitumor therapy.

KEYWORDS gastric cancer, PD-1, PD-L1, stroma, prognosis.

INTRODUCTION

Gastric cancer (GC) is one of the most common cancers worldwide and one of the major causes of mortality. The incidence of cancer is higher in men than it is in women [1]. A large number of different factors, including *Helicobacter pylori* infection [2], smoking [3], dietary habits [2], genetic disorders [4], and others, lead to the appearance of GC. Although the majority of etiological factors of GC appearance are known, early diagnosis of the disease remains somewhat challenging due to its asymptomatic development, and, more often than not, the pathology is diagnosed at late stages. Combination regimens, including fluoropyrimidine and platinum drugs (and trastuzumab in the cases of HER2-positive tumors) in the first line and paclitaxel with or without ramucirumab, in the second line, are standard treatments for advanced GC. However, the median survival time in advanced GC remains approximately 12 to 15 months, obviously as we await new therapies to come on line [5–7]. Immune checkpoint inhibitors (ICIs) have recently become the new standard treatment for several malignancies, including advanced cancer. However, the success currently enjoyed with immunotherapy for GC remains limited. There are several clinical trials focusing on different combinations of immunotherapy and chemotherapy drugs to maximize efficacy. It also remains controversial whether the number of PD-L1-positive tumor cells affects the effectiveness of therapy and whether their number should be considered when prescribing an appropriate treatment. In addition, the qualitative and quantitative composition of the tumor microenvironment can affect the success of GC therapy. For example, an increased number of Th1 cells promotes inflammation and the development of cancer [8], and the content of B cells expressing IL-10 affects the production of cytokines by CD4+ and CD8+ T cells [9].

The main types of tumor immune infiltrate cells include macrophages and T-cells, as well as B-cells. It is known that the number of stromal cells and their population composition may be a prognostic factor for both the course of the disease and response to therapy. PU.1 is a transcription factor that plays an important role in hematopoiesis, and its expression at a high level is characteristic of macrophages. We have previously shown that, for various types of solid tumors, PU.1 can be used as a marker of tumor-associated macrophages [10]. CD3 is a surface marker of mature T cells and is used to determine their total content in various tissue types. CD20 is a transmembrane protein expressed on the surface of B-cell precursors and mature B-cells, allowing its use in various clinical studies as a general B-cell marker.

The purpose of this work was to perform a comprehensive analysis of PD-L1 expression in tumor and stromal GC cells, as well as the content of the soluble form of PD-L1 in the blood plasma of patients. In addition, we analyzed the content of tumor-associated macrophages and B-cells in the stroma of GC tumors.

EXPERIMENTAL

The study included 63 primary GC patients at different stages of the tumor process and 60 healthy donors who underwent examination and treatment at the N.N. Blokhin National Medical Research Center for Oncology of the Ministry of Health of Russia. All procedures performed in the study involving patients and healthy donors met the ethical standards of the organization's ethics committee and the 1964 Declaration of Helsinki and its subsequent amendments or comparable ethical standards. Informed consent was obtained from each of the participants included in the study. The clinical diagnosis of gastric cancer in all patients was confirmed by a morphological examination of the tumor according to the International Histological Classification of Tumors of the Digestive System (WHO, 2019). A description of the studied sample is presented in *Table 1*.

Table 1. Clinical and morphological characteristics of patients with gastric cancer

Characteristics	Number of cases, %
Age	
⩽ 61	32 (51)
> 61	31 (49)
Gender	
Male	35 (56)
Female	28 (44)
Histology	
Adenocarcinoma	52 (82.5)
Signet-ring cell carcinoma	10 (16)
Undifferentiated cancer	1 (1.5)
Stage	
I–II	25 (40)
III–IV	38 (60)
Localization	
Distal	14 (22)
CEC (cardioesophageal cancer)	3 (5)
Proximal	16 (25)
Stomach body	26 (42)
Total lesion	4 (6)
Tumor size (T)	
T1–T2	13 (21)
T3–T4	50 (79)
Nodal status (N)	
N0	24 (38)
N+	39 (62)
Metastasis (M)	
M0	54 (86)
M+	9 (14)
Grade (G)	
G1–G2	19 (30)
G3	44 (70)

The concentration of sPD-L1 and sPD-1 proteins was determined in blood plasma obtained according to the standard technique before specific treatment using Human PD-L1 Platinum ELISA and Human PD-1 ELISA kits (Affimetrix, eBioscience, USA) according to the manufacturer's instructions. Measurements were performed on a BEP 2000 Advance automated enzyme immunoassay (Siemens Healthcare Diagnostics, Germany). The protein content was expressed in picograms (pg) per 1 ml of blood plasma.

Immunohistochemical (IHC)-study of CD163, PU.1, and CD20 was performed according to the standard technique on tumor tissue sections. Tris-EDTA buffer pH 9.0 (PrimeBioMed, Russia) was used for antigen retrieval. The primary antibodies to PU.1 (4G6; PrimeBioMed, Russia, dilution 1 : 200), CD163 (10D6; BIOCARE, USA, dilution 1 : 100), and CD20 (clone PBM-12F1; PrimeBioMed, Moscow, dilution 1 : 100) were incubated for 30 min. The PrimeVision Ms/Rb HRP/DAB detection system (78-310004, PrimeBioMed, Russia) was used according to the manufacturer's instructions.

The preparations obtained were evaluated using an OLYMPUS BX53 microscope, a Lumenera INFINITY2-2C camera, and the Infinity analyze software. The expression of CD163, PU.1, and CD20 was assessed in the tumor stroma. In each case, the number of CD163-, PU.1-, and CD20-positive cells was an-

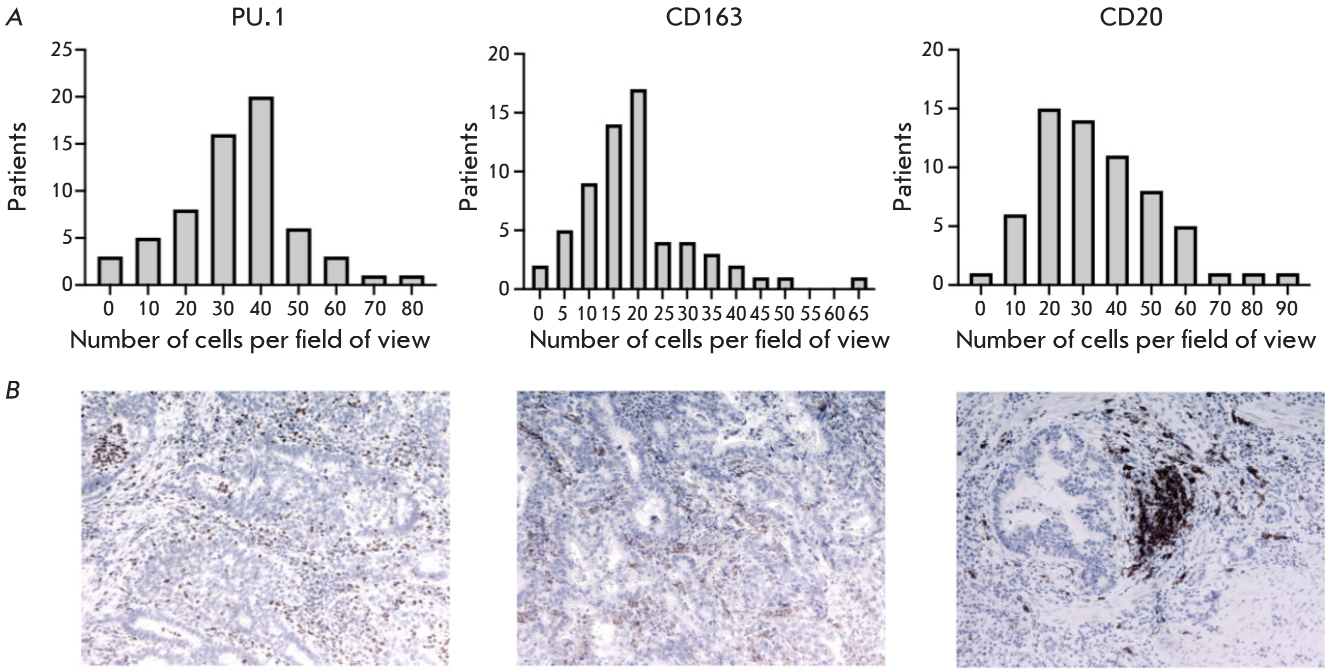


Fig. 1. (*A*) Distribution of PU.1, CD163, and CD20 in the stroma of the tumors of GC patients. (*B*) Immunohistochemical staining of gastric tumors using antibodies to PU.1, CD163, and CD20 (×100)

alyzed under ×200 magnification in five independent fields of view by direct counting. The sample was considered positive if at least one specifically stained cell was present. The content of CD163, PU.1, and CD20 in the tumor stroma was expressed as the average number of cells per field of view.

The data obtained were processed using the GraphPad Prizm 9.0 software. Mann-Whitney nonparametric test and Spearman rank correlation coefficient were used to compare the parameters and analyze their relationships. For the overall survival rate analysis, the patients were divided into two comparison groups depending on the median content of the studied proteins. The analysis of overall survival was performed by constructing survival curves according to the Kaplan-Meier method. The comparison of the statistical significance of differences was performed using the logarithmic rank criterion. To assess the potential impact of various risk factors on survival, we additionally performed a multivariate analysis using a nonparametric Cox proportional hazards model. Differences and correlations were considered statistically significant at $p < 0.05$.

RESULTS

Expression of PU.1, CD163, and CD20 was detected in 100% of the examined GC samples. The distribution of cell numbers in the GC samples is shown in *Fig. 1*.

The analysis of the results showed that the median number of PU.1+ cells in the sample was 34.8 (0.4–77.8) cells per field of view, CD163+ cells – 17.6 (0.8–66.4), CD20+ cells – 32.2 (3.2–91.2). It should be noted that, in gastric tumors, B-cells are present in similar numbers as PU.1+ macrophages.

Association of PU.1, CD163, and CD20 content with clinical and morphological characteristics of GC

At the next stage, we analyzed the association of the PU.1+, CD20+, and CD163+ cell content in the tumor stroma with the clinical and morphological characteristics of the disease (*Table 2*).

The analysis showed that the PU.1 content was significantly associated with tumor size; i.e., larger tumors were characterized by a smaller number of PU.1+ infiltrating cells. We should also note the differences in the content of PU.1+ and CD163+ cells, depending on tumor localization. Thus, in the case of a total gastric lesion, the highest number of PU.1+ cells and the lowest number of CD163+ cells were observed. But these observations did not reach the threshold of statistical significance.

PD-1 and PD-L1 content in tumor samples of GC patients

In addition to analyzing the expression of stromal markers, we assessed the tissue content of PD-L1 in

Table 2. Association of the PU.1+, CD163+, and CD20+ cell content in the tumor stroma with the clinical and morphological characteristics of the disease

Characteristics	PU.1 (number of cells)		CD163 (number of cells)		CD20 (number of cells)	
	Median (25–75%)	p	Median (25–75%)	p	Median (25–75%)	p
Age						
⩽ 61	35.8 (23.4–42.7)	0.488	17.2 (9.05–22.3)	0.297	28.2 (19.4–45.4)	0.418
> 61	34.2 (20.2–42.0)		18.2 (13.2–25.2)		34.4 (20.8–45.2)	
Gender						
Male	33.6 (20.2–37.6)	0.150	16.2 (10.4–24.4)	0.713	29.4 (18.8–45.2)	0.403
Female	37.3 (26.9–44.2)		18.0 (12.9–21.1)		33.9 (22.9–44.9)	
Histology						
Adenocarcinoma	35.0 (20.8–41.9)	0.216	17.6 (11.8–24.1)	0.459	33.3 (19.9–44.8)	0.574
Signet ring cell carcinoma	29.7 (23.3–37.7)		17.5 (10.8–19.7)		30.5 (22.1–47.8)	
Undifferentiated cancer	63.4 (63.4–63.4)		28.2 (28.2–28.2)		19.6 (19.6–19.6)	
Stage						
I–II	35.8 (27.5–44.8)	0.249	17.8 (13.7–20.7)	0.623	34.4 (21.6–46.3)	0.424
III–IV	33.8 (18.0–39.5)		16.7 (10.1–24.6)		29.4 (19.5–43.9)	
Localization						
Distal	34.8 (28.9–44.1)	0.226	17.5 (12.4–23.4)	0.824	33.8 (24.7–43.1)	0.316
CEC (cardioesophageal cancer)	35.8 (0.4–41.8)		19.4 (14.8–25.2)		23.4 (6.0–24.6)	
Proximal	33.6 (12.5–41.2)		17.3 (11.8–24.1)		28.8 (13.3–52.4)	
Stomach body	33.4 (24.4–39.8)		18.1 (11.6–23.6)		37.7 (22.3–47.3)	
Total lesion	48.4 (38.1–61.0)		11.1 (7.7–24.5)		26.5 (18.6–41.1)	
Tumor size (T)						
T1–T2	41.8 (35.5–54.4)	0.005*	17.8 (14.5–23.0)	0.504	36.0 (26.1–48.1)	0.277
T3–T4	32.9 (19.1–38.7)		17.6 (10.1–23.4)		29.4 (19.5–43.9)	
Nodal status (N)						
N0	35.5 (25.3–42.8)	0.733	17.3 (12.9–20.0)	0.437	33.3 (19.9–44.4)	0.947
N+	34.8 (20.2–41.4)		17.8 (11.4–28.2)		29.4 (20.6–45.2)	
Metastasis (M)						
M0	34.8 (25.6–41.9)	0.889	17.6 (11.2–23.4)	0.598	33.3 (21.1–45.3)	0.214
M+	35.4 (13.6–48.6)		18.4 (12.8–25.4)		29.4 (10.6–40.2)	
Grade (G)						
G1–G2	37.6 (25.0–49.8)	0.131	19.0 (13.2–25.8)	0.448	33.2 (22.0–43.6)	0.796
G3–G4	34.2 (18.7–38.9)		16.2 (10.9–21.5)		33.4 (18.6–45.4)	

* Statistically significant.

the studied GC samples. Examples of immunohistochemical staining for PD-L1 are shown in *Fig 2*.

PD-L1 expression in the tumor cells was detected in 35% (22 of 63) of the samples. PD-L1 expression in stromal cells was detected in 60% (38 of 63) of the samples. Then, we analyzed the association of the PD-L1 content with the clinical and morphological characteristics of the disease (*Table 3*).

This study showed that the PD-L1 content in tumor cells had no association with the clinical and morphological characteristics of GC. PD-L1 expression in tumor stromal cells was found to be associated with the presence of distant metastases; i.e., PD-L1 expression in the primary tumor stroma was observed less frequently in their presence.

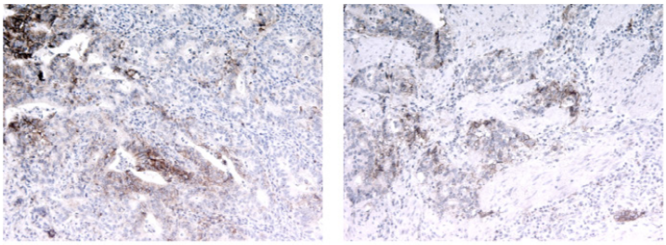


Fig. 2. PD-L1 expression in GC samples (×100)

Soluble forms of PD-1 and PD-L1

In addition, we analyzed the content of soluble forms of the proteins (sPD-1, sPD-L1) of the immunity checkpoint PD-1/PD-L1 in the plasma of RC patients in order to attempt to identify any correlations between their content in plasma and tissue expression and prognostic significance.

At the first stage, we assessed the diagnostic potential of the studied proteins. The median sPD-1 and sPD-L1 content in the blood plasma of

Table 3. Association of PD-L1 content in tumor cells and tumor stroma with the clinical and morphological characteristics of the disease

Characteristics	PD-L1 tumor (*n*)			PD-L1 stroma (*n*)		
	+	-	*p*	+	-	*p*
Age ⩽ 61 > 61	8 14	24 17	0.117	18 20	14 11	0.609
Gender Male Female	12 10	23 18	> 0.999	18 20	17 8	0.127
Histology Adenocarcinoma Signet ring cell carcinoma Undifferentiated cancer	19 3 0	33 7 1	0.704	32 5 1	20 5 0	0.567
Stage I–II III–IV	8 14	17 24	0.790	16 22	9 15	0.793
Localization Distal CEC (cardioesophageal cancer) Proximal Stomach body Total lesion	3 0 7 11 1	11 3 9 15 3	0.396	8 2 10 16 2	6 1 6 10 2	0.987
Tumor size (T) T1–T2 T3–T4	3 19	10 31	0.515	11 27	2 23	0.058
Nodal status (N) N0 N+	7 15	17 24	0.588	13 25	11 14	0.597
Metastasis (M) M0 M+	21 1	33 8	0.144	36 2	18 7	0.023*
Grade (G) G1–G2 G3	7 12	12 21	> 0.999	14 18	5 15	0.239

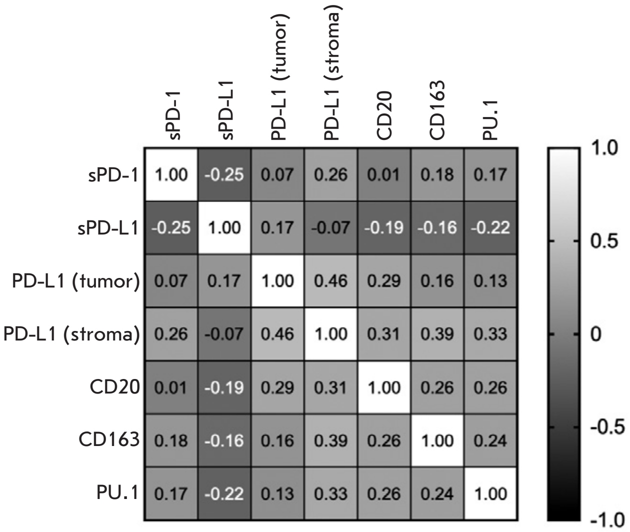
*Statistically significant.

healthy donors was 29.25 (14.9–45.5) pg/ml and 36.23 (9.83–73.1) pg/ml, respectively; and in the group of GC patients – 12.57 (7.7–19.7) pg/ml and 21.83 (10.1–74.3) pg/ml. The statistical analysis showed that the content of the soluble form of the sPD-1 receptor was significantly lower in GC patients compared to the healthy donors. The levels of sPD-L1 did not differ between the groups of healthy donors and GC patients.

Correlation analysis of soluble and tissue forms of the studied proteins

We performed a correlation analysis of the proteins examined by determining the Spearman rank correlation coefficient. The results are shown in *Fig. 3*.

The analysis showed that the plasma content of the soluble form of the sPD-1 receptor inversely correlates with the plasma content of sPD-L1 and directly correlates with the tissue expression of PD-L1 in stromal cells (r = -0.251; p = 0.047 and r = 0.255;



	sPD-1	sPD-L1	PD-L1 (tumor)	PD-L1 (stroma)	CD20	CD163	PU.1
sPD-1	1.00	-0.25	0.07	0.26	0.01	0.18	0.17
sPD-L1	-0.25	1.00	0.17	-0.07	-0.19	-0.16	-0.22
PD-L1 (tumor)	0.07	0.17	1.00	0.46	0.29	0.16	0.13
PD-L1 (stroma)	0.26	-0.07	0.46	1.00	0.31	0.39	0.33
CD20	0.01	-0.19	0.29	0.31	1.00	0.26	0.26
CD163	0.18	-0.16	0.16	0.39	0.26	1.00	0.24
PU.1	0.17	-0.22	0.13	0.33	0.26	0.24	1.00

Fig. 3. Correlation analysis between tissue and serum levels of PD-1, PD-L1, PU.1, CD163, and CD20 in gastric cancer patients

$p = 0.044$, respectively). Also, PD-L1 expression in the stromal cells of gastric tumors directly correlates with PD-L1 expression in tumor cells and the content of all stromal markers examined. A similar pattern was observed for B cells: namely, the content of CD20+ cells in tumor stroma positively correlates with both macrophage content and PD-L1 expression in both stroma and tumor cells, and this correlation was statistically significant.

Prognostic significance of PD-L1/PD-1 in cancer patients

We analyzed the prognostic significance of the markers studied and their combinations in GC patients. Depending on the content of the soluble forms of the studied proteins, patients were divided into two groups: those with a high and low content of the studied markers relative to the median. In the case of the PD-L1 tissue expression, patients were divided into two groups: depending on the presence or absence of this protein separately in tumor and stromal cells. In addition, we analyzed survival depending on the complex content of both soluble sPD-L1 and the tissue form of PD-L1. The survival plots of patients are shown in *Fig. 4*.

This study failed to establish a relationship between the sPD-1 and sPD-L1 levels in GC patients and the survival prognosis. For the tissue form of PD-L1, an inconsistent pattern was revealed. However, it should be noted that, for PD-L1 in tumor cells, we observed a trend toward the prognostic significance of the marker; i.e., a high expression of this protein in tumor cells of GC patients is a more favorable prognostic factor than a low expression of the marker ($p = 0.122$). Also, a comprehensive analysis indicated that a simultaneous high content of tissue and soluble forms of PD-L1 was not a prognostic marker in cancer.

Next, we analyzed the prognostic significance of PU.1, CD20, and CD163 in cancer. The results are shown in *Fig. 5*.

The data in *Fig. 4* show that the studied stromal markers (PU.1, CD163, and CD20) are not prognostically significant in GC.

In addition, we performed a multivariate statistical analysis of the prognostic significance of all investigated markers. The results are presented in *Table 4*.

Cox regression analysis revealed that a high CD163 content in cancer is an independent prognostic factor associated with decreased overall survival.

DISCUSSION

The clinical and prognostic significance of the microenvironment of gastric tumors is being actively studied. In this work, we analyzed the content of PU.1+, CD163+, and CD20+ in the stroma of gastric tumors and evaluated their clinical and prognostic significance. In the context of solid tumors, the clinical significance of PU.1 expression was studied in patients with breast cancer and gliomas [11, 12]. Association of its expression with progression of the disease and an unfavorable prognosis were established for both tumor types. PU.1 expression has also been studied in non-small cell lung cancer (NSCLC) [13], colorectal cancer [14], and esophageal cancer [10]. One study was devoted to the study of PU.1 expression in GC, which showed that PU.1 expression is significantly elevated in gastric tumor tissue compared to the relative norm and is associated with an unfavorable prognosis and disease progression. Moreover, high PU.1 expression positively correlates with the number of activated CD4 memory T cells, resting NK cells, M2 macrophages, resting dendritic cells, and neutrophils in the tumor stroma [15]. Our study failed to reveal any prognostic significance of this protein, but consistent with the literature data, we observed a positive correlation of the PU.1+ cell content with macrophages and B-cells, as well as PD-L1+ cells in the tumor stroma.

A large number of studies are devoted to the analysis of the CD163+ macrophage content in gastric tumors, but the results are rather inconclusive. The literature suggests that CD163 expression is often associated with an unfavorable prognosis of various solid tumors [16]. However, for gastrointestinal tumors, it has been shown that CD163 can be a marker of good prognosis, particularly in esophageal cancer [17] and colorectal cancer [18]. For GC, an increased density of CD163+ macrophages in tumor stroma has been shown to be associated with the activation of the immune response and improved patient survival according to single-factor analysis [19]. However, opposite results have also been reported. A study of 148 tumor tissue samples revealed that high CD68+/CD163+ infiltration was a marker of unfavorable prognosis [20]. Other researchers demonstrated that an elevated CD163+ cell content was associated with large tumor size, low tumor differentiation, and metastases in regional lymph nodes. Moreover, the CD163 density increased with the depth of invasion, stage of the disease, and increased expression of tumor stem cell markers. The authors also found that an increased expression of this marker was associated with disease recurrence [21, 22]. The data we obtained are in agreement with the literature data; namely, a high content of CD163+ cells in the tumors of GC patients is an independent marker of an unfavorable prognosis in this pathology. There is also evidence in the litera-

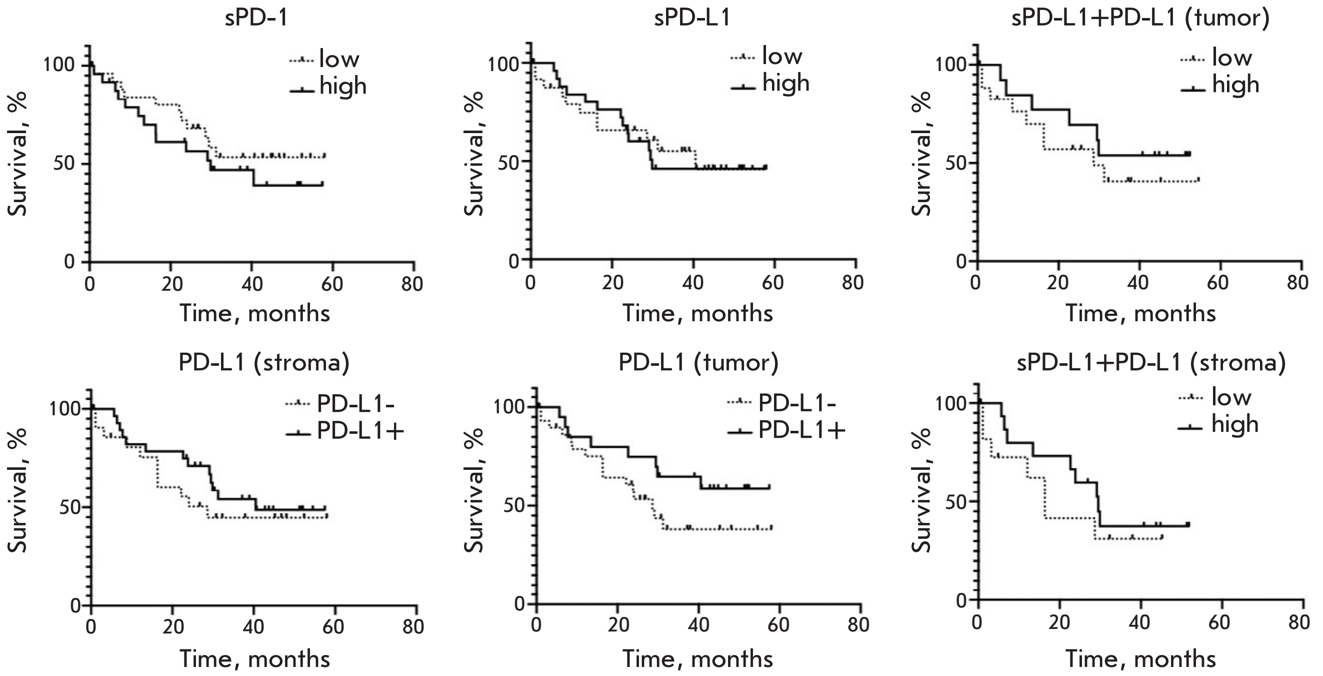


Fig. 4. Analysis of the overall survival of GC patients depending on the content of soluble (sPD-L1, sPD-1) and tissue (PD-L1, PD-1) forms of the main components of the PD-1/PD-L1 immunity checkpoint

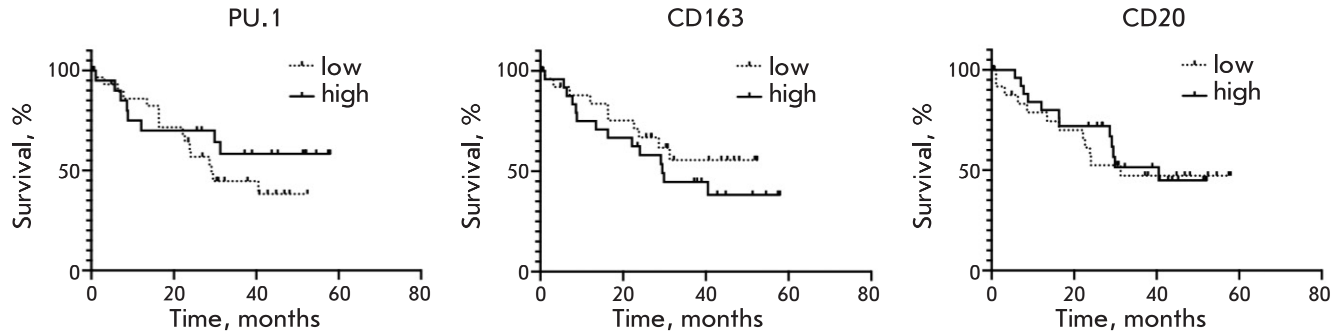


Fig. 5. Analysis of the overall survival of GC patients depending on the PU.1, CD163, and CD20 content in the tumor stroma

Table 4. Statistical analysis of the prognostic significance of sPD-1, sPD-L1, PD-1, PD-L1, CD20, CD163, and PU.1 in GC patients

Metrics	Univariate analysis			Multivariate analysis		
	HR	95% CI	*p*	HR	95% CI	*p*
sPD-1 (high/low)	1.443	(0.646–3.226)	0.366	0.971	(0.915–1.013)	0.234
sPD-L1 (high/low)	1.038	(0.466–2.315)	0.927	0.999	(0.988–1.008)	0.780
PD-L1 (tumor) (high/low)	0.524	(0.235–1.167)	0.122	0.480	(0.150–1.406)	0.193
PD-L1 (stroma) (high/low)	0.721	(0.316–1.644)	0.419	0.954	(0.332–2.564)	0.927
CD20 (high/low)	0.876	(0.393–1.953)	0.745	0.992	(0.965–1.016)	0.526
CD163 (high/low)	1.509	(0.677–3.361)	0.316	1.053	(1.007–1.098)	0.019*
PU.1 (high/low)	0.654	(0.292–1.466)	0.319	0.991	(0.963–1.018)	0.497

*Statistically significant.

ture that increased CD163 expression is characteristic of PD-L1+ cancer compared to PD-L1 [23]. Our results demonstrate that the CD163+ cell content in tumor stroma positively correlates with PD-L1 expression in stromal but not in tumor cells in GC.

At the next stage of the study, we analyzed the content of CD20+ cells in the stroma of the tumors of GC patients. Various studies report the presence of CD20+ B-lymphocytes in tumors of different types to have an ambiguous effect on survival prognosis and tumor stage [24]. For example, it was shown that in breast cancer, the total number of CD20+ B-lymphocytes is associated with tumor progression [25], while in some cases of ovarian, liver, and colorectal cancer, the correlation was the inverse [26–28]. The increased content of CD20+ B-lymphocytes in the stroma was shown to be associated with a better prognosis for GC patients. However, no association between the B-lymphocyte count and clinical and morphological characteristics was revealed [29]. Other researchers have demonstrated similar results, showing that a higher CD20+ B-cell density in the stroma is associated with a better prognosis. This study has also found the CD20 expression to be associated with CD68 in the tumor stroma. Interestingly, some stromal immune cells expressed Ki-67 and these were mostly CD20+ cells. Moreover, a combination of Ki-67+ and CD20+ demonstrated better prognostic potential for GC [30]. The results of our study demonstrate the lack of prognostic significance of CD20 in GC, indicating the need to use combinations of markers to improve the effectiveness of predicting the clinical course of the disease.

About two dozen studies are devoted to the prognostic significance of PD-L1 tissue expression. Most of those studies suggest an unfavorable prognostic significance of this protein expression in GC tumor cells [31]. However, some studies suggest high PD-L1 expression in tumor cells to be a good prognosis marker [32, 33]. Our study has demonstrated that PD-L1 expression in tumor cells is associated with a higher overall survival chance for patients, with no such pattern found for PD-L1 expression in stromal cells or the concentration of its soluble form in plasma.

CONCLUSION

The results obtained in this study suggest that markers of stromal cells in gastric malignancies can potentially be used to plot treatment strategies and disease prognosis. However, current techniques, namely single-color immunohistochemistry, do not provide a sufficiently informative response. In order to use stromal markers effectively in the case of GC, the development of a comprehensive assay involving the determination of several serum markers and a multiplex analysis of several tumor stroma markers is needed. ●

Conflict of interest: The authors declare no conflict of interest.

The study was supported by the Russian Foundation for Basic Research (project No. 20-015-00479).

REFERENCES

1. Bray F., Ferlay J., Soerjomataram I., Siegel R.L., Torre L.A., Jemal A. // CA Cancer J. Clin. 2018. V. 68. № 6. P. 394–424.
2. Gonzalez C.A., Sala N., Rokkas T. // Helicobacter. 2013. V. 18 Suppl 1. P. 34–38.
3. Nomura A., Grove J.S., Stemmermann G.N., Severson R.K. // Cancer Res. 1990. V. 50. № 21. P. 7084.
4. Brooks-Wilson A.R., Kaurah P., Suriano G., Leach S., Senz J., Grehan N., Butterfield Y.S., Jeyes J., Schinas J., Bacani J., et al. // J. Med. Genet. 2004. V. 41. № 7. P. 508–517.
5. Cunningham D., Starling N., Rao S., Iveson T., Nicolson M., Coxon F., Middleton G., Daniel F., Oates J., Norman A.R., et al. // N. Engl. J. Med. 2008. V. 358. № 1. P. 36–46.
6. Bang Y.J., van Cutsem E., Feyereislova A., Chung H.C., Shen L., Sawaki A., Lordick F., Ohtsu A., Omuro Y., Satoh T., et al. // Lancet. 2010. V. 376. № 9742. P. 687–697.
7. Wilke H., Muro K., van Cutsem E., Oh S.C., Bodoky G., Shimada Y., Hironaka S., Sugimoto N., Lipatov O., Kim T.Y., et al. // Lancet Oncol. 2014. V. 15. № 11. P. 1224–1235.
8. Zhang H., Yue R., Zhao P., Yu X., Li J., Ma G., Tang J., Zhang L., Feng L., Sun L., et al. // Tumour Biol. 2017. V. 39. № 6. P. 1010428317705747. doi: 10.1177/1010428317705747
9. Hu H.T., Ai X., Lu M., Song Z., Li H. // Exp. Cell. Res. 2019. V. 384. № 2. P. 111652.
10. Kovaleva O.V., Rashidova M.A., Samoilova D.V., Podlesnaya P.A., Mochalnikova V.V., Gratchev A. // Anal. Cell. Pathol. (Amst.). 2020. V. 2020. P. 5424780.
11. Xu Y., Gu S., Bi Y., Qi X., Yan Y., Lou M. // Oncol. Lett. 2018. V. 15. № 3. P. 3753–3759.
12. Lin J., Liu W., Luan T., Yuan L., Jiang W., Cai H., Yuan W., Wang Y., Zhang Q., Wang L. // Oncol. Lett. 2017. V. 14. № 6. P. 8220–8226.
13. Kovaleva O.V., Rashidova M.A., Samoilova D.V., Podlesnaya P.A., Mochalnikova V.V., Gratchev A.N. // Bull. Exp. Biol. Med. 2021. V. 170. № 4. P. 489–492.
14. Kovaleva O.V., Gratchev A.N., Podlesnaya P.A., Rashidova M.A., Samoilova D.V., Sokolov N Yu., Mamedli Z.Z., Kudlay D.A., Kushlinskii N.E. // Clinical and experimantal morphology. 2021. V. 10. № 2. P. 32–39. (In Russ.)
15. Huang J., Chen W., Jie Z., Jiang M. // Front. Oncol. 2022. V. 12. P. 820568.
16. Mantovani A., Sozzani S., Locati M., Allavena P., Sica A. // Trends Immunol. 2002. V. 23. № 11. P. 549–555.

17. Kovaleva O., Podlesnaya P., Rashidova M., Samoilova D., Petrenko A., Mochalnikova V., Kataev V., Khlopko Y., Plotnikov A., Gratchev A. // Biomedicines. 2021. V. 9. № 7. P. 743.
18. Koelzer V.H., Canonica K., Dawson H., Sokol L., Karamitopoulou-Diamantis E., Lugli A., Zlobec I. // Oncoimmunology. 2016. V. 5. № 4. P. e1106677.
19. Huang Y.K., Wang M., Sun Y., Di Costanzo N., Mitchell C., Achuthan A., Hamilton J.A., Busuttil R.A., Boussioutas A. // Nat. Commun. 2019. V. 10. № 1. P. 3928.
20. Svensson M.C., Svensson M., Nodin B., Borg D., Hedner C., Hjalmarsson C., Leandersson K., Jirstrom K. // J. Innate Immun. 2022. № 3. P. 1–14.
21. Zhu Q., Wu X., Tang M., Wu L. // Medicine (Baltimore). 2020. V. 99. № 17. P. e19839.
22. Zhang W.J., Zhou Z.H., Guo M., Yang L.Q., Xu Y.Y., Pang T.H., Gao S.T., Xu X.Y., Sun Q., Feng M., et al. // J. Cancer. 2017. V. 8. № 3. P. 363–370.
23. Harada K., Dong X., Estrella J.S., Correa A.M., Xu Y., Hofstetter W.L., Sudo K., Onodera H., Suzuki K., Suzuki A., et al. // Gastric Cancer. 2018. V. 21. № 1. P. 31–40.
24. Sjoberg E., Frodin M., Lovrot J., Mezheyeuski A., Johansson M., Harmenberg U., Egevad L., Sandstrom P., Ostman A. // Br. J. Cancer. 2018. V. 119. № 7. P. 840–846.
25. Mahmoud S.M., Lee A.H., Paish E.C., Macmillan R.D., Ellis I.O., Green A.R. // Breast Cancer Res. Treat. 2012. V. 132. № 2. P. 545–553.
26. Shi J.Y., Gao Q., Wang Z.C., Zhou J., Wang X.Y., Min Z.H., Shi Y.H., Shi G.M., Ding Z.B., Ke A.W., et al. // Clin. Cancer Res. 2013. V. 19. № 21. P. 5994–6005.
27. Lundgren S., Berntsson J., Nodin B., Micke P., Jirstrom K. // J. Ovarian Res. 2016. V. 9. P. 21.
28. Berntsson J., Nodin B., Eberhard J., Micke P., Jirstrom K. // Int. J. Cancer. 2016. V. 139. № 5. P. 1129–1139.
29. Dong J., Li J., Liu S.M., Feng X.Y., Chen S., Chen Y.B., Zhang X.S. // Med. Oncol. 2013. V. 30. № 1. P. 442.
30. Meier A., Nekolla K., Hewitt L.C., Earle S., Yoshikawa T., Oshima T., Miyagi Y., Huss R., Schmidt G., Grabsch H.I. // J. Pathol. Clin. Res. 2020. V. 6. № 4. P. 273–282.
31. Gu L., Chen M., Guo D., Zhu H., Zhang W., Pan J., Zhong X., Li X., Qian H., Wang X. // PLoS One. 2017. V. 12. № 8. P. e0182692.
32. Kim J.W., Nam K.H., Ahn S.H., Park D.J., Kim H.H., Kim S.H., Chang H., Lee J.O., Kim Y.J., Lee H.S., et al. // Gastric Cancer. 2016. V. 19. № 1. P. 42–52.
33. Boger C., Behrens H.M., Mathiak M., Kruger S., Kalthoff H., Rocken C. // Oncotarget. 2016. V. 7. № 17. P. 24269–24283.

Multiple Sclerosis Is Associated with Immunoglobulin Germline Gene Variation of Transitional B Cells

Y. A. Lomakin[1*], L. A. Ovchinnikova[1], M. N. Zakharova[2], M. V. Ivanova[2], T. O. Simaniv[2], M. R. Kabilov[3], N. A. Bykova[4], V. S. Mukhina[4,5], A. N. Kaminskaya[1], A. E. Tupikin[3], M. Y. Zakharova[1], A. V. Favorov[4], S. N. Illarioshkin[2], A. A. Belogurov[1,6], A. G. Gabibov[1,7]

[1]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, 117997 Russia
[2]Research Center of Neurology, Moscow, 125367 Russia
[3]Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, 630090 Russia
[4]Vavilov Institute of General Genetics RAS, Moscow, 119991 Russia
[5]Institute for information transmission problems RAS, Moscow, 127051 Russia
[6]A.I. Yevdokimov Moscow State University of Medicine and Dentistry, Moscow, 127473 Russia
[7]Lomonosov Moscow State University, Moscow, 119991 Russia
[*]E-mail: lomakin@ibch.ru

Received September 09, 2022; in final form, October 20, 2022
DOI: 10.32607/actanaturae.11794

ABSTRACT The regulatory functions of the B-cell compartment play an important role in the development and suppression of the immune response. Disruption of their anti-inflammatory functions may lead to the acceleration of immunopathological processes, and to autoimmune diseases, in particular. Unfortunately, the exact mechanism underlying the functioning and development of regulatory B cells (Breg) has not yet been fully elucidated. Almost nothing is known about their specificity and the structure of their B-cell receptors (BCRs). In this research, we analyzed the BCR repertoire of the transitional Breg (tBreg) subpopulation with the CD19^{+}CD24^{high}CD38^{high} phenotype in patients with multiple sclerosis (MS), using next-generation sequencing (NGS). We show, for the first time, that the immunoglobulin germline distribution in the tBreg subpopulation is different between MS patients and healthy donors. The registered variation was more significant in patients with a more severe form of the disease, highly active MS (HAMS), compared to those with benign MS (BMS). Our data suggest that during MS development, deviations in the immunoglobulin Breg repertoire occur already at the early stage of B-cell maturation, namely at the stage of tBregs: between immature B cells in the bone marrow and mature peripheral B cells.

KEYWORDS multiple sclerosis, neurodegeneration, immunoglobulin, transitional B cells, NGS, regulatory B cells, BCR-Seq, germline, TrB.

ABBREVIATIONS MS – multiple sclerosis; CNS – central nervous system; HAMS – highly active multiple sclerosis; BMS – benign multiple sclerosis; Breg – regulatory B cell; tBreg – transitional regulatory B cell.

INTRODUCTION

Multiple sclerosis (MS) is one of the most common chronic autoimmune diseases of the central nervous system (CNS). It affects more than 2.3 million people worldwide [1]. Its triggering mechanism, and its mechanism of immune-mediated neurodegeneration in particular, still remains unknown, which causes significant difficulties in efforts to design a strategy for MS treatment and drug development [2–4]. Since MS was discovered, the main role in its pathogenesis has been assigned exclusively to the T cell-mediated immunity. However, over the past decade, a lot of evidence has emerged confirming that B cells are directly involved in the development of autoimmune processes, including MS [5]. MS patients have been shown to have elevated titers of autoreactive antibodies that specifically recognize native components of the myelin sheath. Moreover, the catalytic immunoglobulins that hydrolyze the myelin basic protein (MBP), one of the characteristic autoantigens in MS, have also been found exceptionally in MS patients, as opposed to healthy donors or patients with other neurodegenerative diseases [6–8]. Despite the long history of MS research, its exact etiology still remains elusive. Molecular mimicry, epitope spreading, and cross-reactivity are believed to underlie the mechanisms of viral induction of the disease [9–15]. The immunoglobulin repertoire of MS patients contains cross-reactive antibodies capable of simultaneously binding the human myelin basic protein and components of the Epstein–Barr virus [15, 16]

Regulatory B cells (Bregs), a new subpopulation of B cells, have recently become the object of increasing attention [17,18]. The fundamental interest in them lies in the need to understand the exact mechanism of suppression of the inflammatory response by B cells. There is no clear understanding at what stage of maturation a B cell acquires regulatory functions and how it is affected by BCR specificity. From a practical point of view, Bregs attract one's attention as the cells directly involved in the development of autoimmune and lymphoproliferative pathologies. However, it is impossible to draw an unambiguous conclusion about the exact deviations that happen during autoimmune inflammation: there can be a change in the number of Breg cells, a disruption of their functions, or a combination of these two phenomena. Furthermore, there is limited information regarding the specificity of Bregs, although it has been shown that they require a B-cell receptor for proper functioning [19]. It is unclear whether the development of an autoimmune response is accompanied by disruptions in the maturation of Breg immunoglobulin genes and whether these regulatory cells can be autoreactive. It is not known at what stage of development the most significant changes in the Breg pool occur: in naïve, transitional, or mature B cells? Earlier, we found an increased number of transitional Bregs (tBregs) in the peripheral blood of MS patients [20]. Notably, the tBregs' immunoglobulin heavy chains in MS patients carry fewer hypermutations compared to healthy donors.

In the present study, we have examined whether the structures of B-cell receptors from one of the most fully described subpopulations of tBregs, $CD19^{+}CD24^{high}CD38^{high}$, differ in MS patients and healthy donors. The NGS analysis of BCR sequences (BCR-Seq) revealed that the distribution of a number of immunoglobulin germline genes in MS patients differs from that in healthy individuals. Moreover, during our analysis of the total pool of B cells and the tBreg subpopulation, both an excess and decrease in germline occurrence were observed in comparison with the normal range. It is important to note that this difference is more pronounced in patients with a more severe disease course (highly active MS (HAMS) [21]) compared to those with benign MS (BMS) [22].

EXPERIMENTAL

MS patients and healthy donors

Peripheral blood was sampled from nine patients with MS and six healthy donors (*Table 1*) at the Sixth Neurological Department of the Research Center of Neurology (Moscow, Russia). The patients with MS were aged 23–61 years (mean, 40.0 ± 9.1). Disease severity according to the EDSS scale ranged from 1.5 to 8.5. EDSS scores from 0 to 10 were calculated using the Kurtzke Expanded Disability Status Scale (EDSS) [23]. Five patients with HAMS and four patients with BMS were selected for the study. Data on the disease course, as well as treatment duration and history, were collected (*Table 1*). The study was approved by the local ethics committee of the Neurology Research Center and was conducted in full compliance with the WMA Declaration of Helsinki, ICH GCP, and relevant local legislation. All patients provided written informed consent after discussion of the study protocol.

Isolation of B cells from peripheral blood

Mononuclear cells from the peripheral blood of MS patients and healthy donors were obtained by sedimentation enrichment using Ficoll density gradient centrifugation. The residual erythrocyte fraction was removed using a ACK lysing buffer. The resulting mononuclear cells were filtered through a 40-mm nylon filter and stained with fluorescent antibodies: α-CD19-PE-Cy7, α-CD24-PE, α-CD38-APC, α-CD45-APC-Cy7 (Bio-legend, USA), and SYTOX Green dead cell stain (ThermoFisher Scientific) for 60 min at +4°C in the dark. The populations of tBregs ($CD19^{+}CD24^{high}CD38^{high}$) and total B cells ($CD19^{+}$) were collected directly into microcentrifuge tubes containing a Qiazol lysis buffer (Qiagen, Germany). Cell sorting was performed using a BD FACSAria III flow cytometer.

Library preparation for immunoglobulin sequencing (RT-PCR)

RNA isolation was performed using an RNeasy Mini Kit (Qiagen, Germany) according to the manufacturer's protocol. Reverse transcription (RT) was carried out using an MMLV RT kit with oligo(dT) and random primers according to the manufacturer's instructions (Evrogen, Russia). Oligonucleotides for the amplification of variable fragments of human immunoglobulins VH and VL contained 15 forward primers for VH and four reverse primers for the human heavy chain J fragment, 13 Vϰ forward primers and two Jϰ reverse primers for the kappa light chain, and 16 Vλ forward primers and three reverse primers Jλ for the lambda light chain [24]. Fifteen VH forward primers were used individually in each sample in a 50 μL reaction mixture with an equimolar mixture of four JH reverse primers. Thirteen Vϰ primers and sixteen Vλ primers were used individually to amplify the VL genes with an appropriate mixture of two Vϰ reverse primers or three Vλ reverse primers in 50 μL of the reaction mixture for each sample. cDNA (0.02 μg) was used as a template in PCR performed with the Hot Start Taq Master Mix kit (Evrogen, Russia). The PCR conditions were as follows: one step (94°C – 3 min); one cycle (94°C – 25 s, 62°C – 25 s, 72°C – 25 s); two cycles (94°C – 25 s, 60°C – 25 s, 72°C – 25 s); two cycles (94°C – 25 s, 58°C – 25 s, 72°C – 25 s); three cycles (94°C – 25 s, 56°C – 25 s, 72°C – 25 s); three cycles (94°C – 25 s, 54°C – 25 s, 72°C – 25 s); 30 cycles (94°C – 25 s, 52°C – 25 s, 72°C – 25 s); and final elongation (72°C – 4 min). PCR mixtures of 15 VH gene samples, 13 Vϰ gene samples, and 16 Vλ gene samples were individually pooled and concentrated to 50–80 μL using an Amicon 30 kDa centrifugal filter unit (Merck, Millipore). The PCR products (~ 400 bp) VH, Vϰ, and Vλ were loaded onto 1.5% agarose gel and purified using an agarose gel DNA purification kit (Monarch, NEB).

Next-generation sequencing of VH, Vϰ, and Vλ variable immunoglobulin fragments

One μg of purified VH, Vϰ, and Vλ PCR product was ligated to NEBNext Multiplex Oligos adapters using the NEBNext Ultra DNA Library Preparation Kit for Illumina (NEB). Libraries were sequenced on a MiSeq system using a 2 × 300 bp sequencing kit (Illumina) at the Genomics Core Facility SB RAS (Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia).

Analysis of the NGS data

The analysis was carried out using the MiXCR software [25] in two stages. Initially, raw sequencing data were processed using the default MiXCR algorithm (align, assemble, export) employing the IMGT library as a germline gene reference. The generated reads successfully aligned with the germline genes and containing the complete immunoglobulin target sequence (CDR1 + FR2 + CDR2 + FR3 + CDR3) were then subjected to resampling to normalize different numbers of reads. When analyzing the occurrence frequency of germline genes, mutations in the variable fragments VH, Vϰ, and Vλ were not taken into account.

Statistical analysis

The statistical analysis was performed using the Prism 6 software utilizing the Mann–Whitney test and paired Student's t-test.

RESULTS AND DISCUSSION

Recently, there has been a growing number of studies demonstrating the importance of B cells in the regulation of autoimmune diseases, including MS [26, 27]. However, the Breg subpopulations in MS patients still have not been fully characterized. To date, very little data have been published on the specificity and structure of their B-cell receptors. For a deeper understanding of the nature of Bregs development and characterization of their maturation, we analyzed the CD19$^+$CD24^{high}CD38^{high} subpopulation, one of the most convincingly-confirmed phenotypic portraits of tBregs, which are at an intermediate stage of development between immature bone marrow cells and fully mature naïve B cells in peripheral blood and secondary lymphoid tissues [28, 29]. Peripheral blood samples were obtained from nine patients with MS and six healthy donors (*Table 1*). Mononuclear cells were stained with antibodies against the CD19, CD24, and CD38 surface markers. The total pool of CD19$^+$ B cells and tBreg CD19$^+$CD24^{high}CD38^{high} were separately obtained by cell sorting for subsequent RNA isolation and sequence analysis of B-cell receptors. For this purpose, the sequences of the variable fragments of the heavy (VH) and light (VL) immunoglobulin chains of each patient were amplified from cDNA synthesized from the isolated RNA, and, then, NGS of the VH, Vϰ, and Vλ genes was performed. For a fair analysis, sequencing of the immunoglobulin repertoire of the total B-cell pool and the subset of transitional Bregs was performed with a read depth of at least five functional reads per sorted cell. After all the stages of bioinformatic filtering, we obtained an average of 83,100 functional sequences for the heavy chain; 37,591 sequences for the kappa chain; and 34,565 sequences for the lambda chain of the total pool of CD19+ cells and tBreg subpopulation with the CD19$^+$CD24^{high}CD38^{high} phenotype. The sequences are available in the ArrayExpress repository (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10859).

To analyze the distribution of the VH, Vk, and Vλ germlines, we used primers capable of amplify-

Table 1. List of MS patients and healthy donors participating in the study

No.	MS phenotype[1]	Age, years	Sex	EDSS[2]	Treatment[3]	Disease duration, years
MS1	BMS	56	female	2.5	No treatment	11
MS2	BMS	61	female	3	No treatment	26
MS3	BMS	43	female	1.5	No treatment	12
MS4	BMS	36	male	2.5	No treatment	14
MS5	HAMS	33	male	6	IFNβ1b (2006–2011; 2014–2017).	12
MS6	HAMS	23	male	5	No treatment	3
MS7	HAMS	37	female	5	IFNβ1b (2014–2016). GA (2016–2017).	5
MS8	HAMS	29	female	8	GA (2012–2014). IVIG (2014). IFNβ1b (2015–2016).	12
MS9	HAMS	39	female	8.5	No treatment	8
HD1	Healthy	24	female	–	–	–
HD2	Healthy	40	female	–	–	–
HD3	Healthy	36	male	–	–	–
HD4	Healthy	27	female	–	–	–
HD5	Healthy	42	female	–	–	–
HD6	Healthy	25	female	–	–	–

[1] BMS – patient with benign MS; HAMS – highly active MS; HD – healthy donor.
[2] EDSS – the Expanded Disability Status Scale.
[3] IFNβ1b – interferon-β-1b; GA – glatiramer acetate; IVIG – intravenous immunoglobulin.

ing almost all possible variants of the V_HDJ_H, V_KJ_K, and $V_λJ_λ$. functional fragments. When comparing the distribution of IgVH genes in patients with MS and healthy donors, all seven functional families of VH were amplified. The IGHV3 germline genes were most abundant in MS patients and healthy donors in both the total B-cell pool and the tBreg subpopulation. The IGHV2-26, IGHV2-5, IGHV2-70 germline immunoglobulin sequences, which are present in small amounts in almost every healthy individual, disappear during the development of MS (*Fig. 1*). In patients with a more severe disease (HAMS), the variation in the repertoire of tBreg immunoglobulin genes increases as compared to healthy donors. The IGHV3-66 germline occurs in healthy donors and patients with BMS at a comparable level, but this gene almost completely disappears in HAMS patients. One of the major germlines, IGHV5-51, is observed in all the analyzed donors, but its frequency decreases significantly in MS patients. Contrariwise, the only IGHV4-31 gene is more frequent in MS patients both in the total pool of B cells and in the tBreg subpopulation. This correlates with the previously published data on increased levels of the IGHV4 family in the B-cell repertoire of the peripheral blood and cerebrospinal fluid of MS patients [30, 31].

The repertoire of germline genes encoding light chains of Breg immunoglobulins also differs between MS patients and healthy donors. The portion of IGKV1-12 germline, which is normally found in approximately 2% of immunoglobulin sequences, decreases below 1% in the case of HAMS patients (*Fig. 2*). In BMS patients with a milder disease course, the frequency of the IGKV1-12 germline does not differ from that in healthy donors. The IGKV1-33 germline is less common not only in HAMS patients, but also in the entire group of MS patients, which includes patients with both disease courses. On the contrary, IGKV2D-24, IGKV3-11, and IGKV6D-21 germline genes are significantly more common in HAMS patients than in healthy donors. The IGKV2D-29, IGKV3D-20, and IGKV6-21 genes are more prevalent both in HAMS patients and in the MS group in general. For the kappa light chain, the distribution of

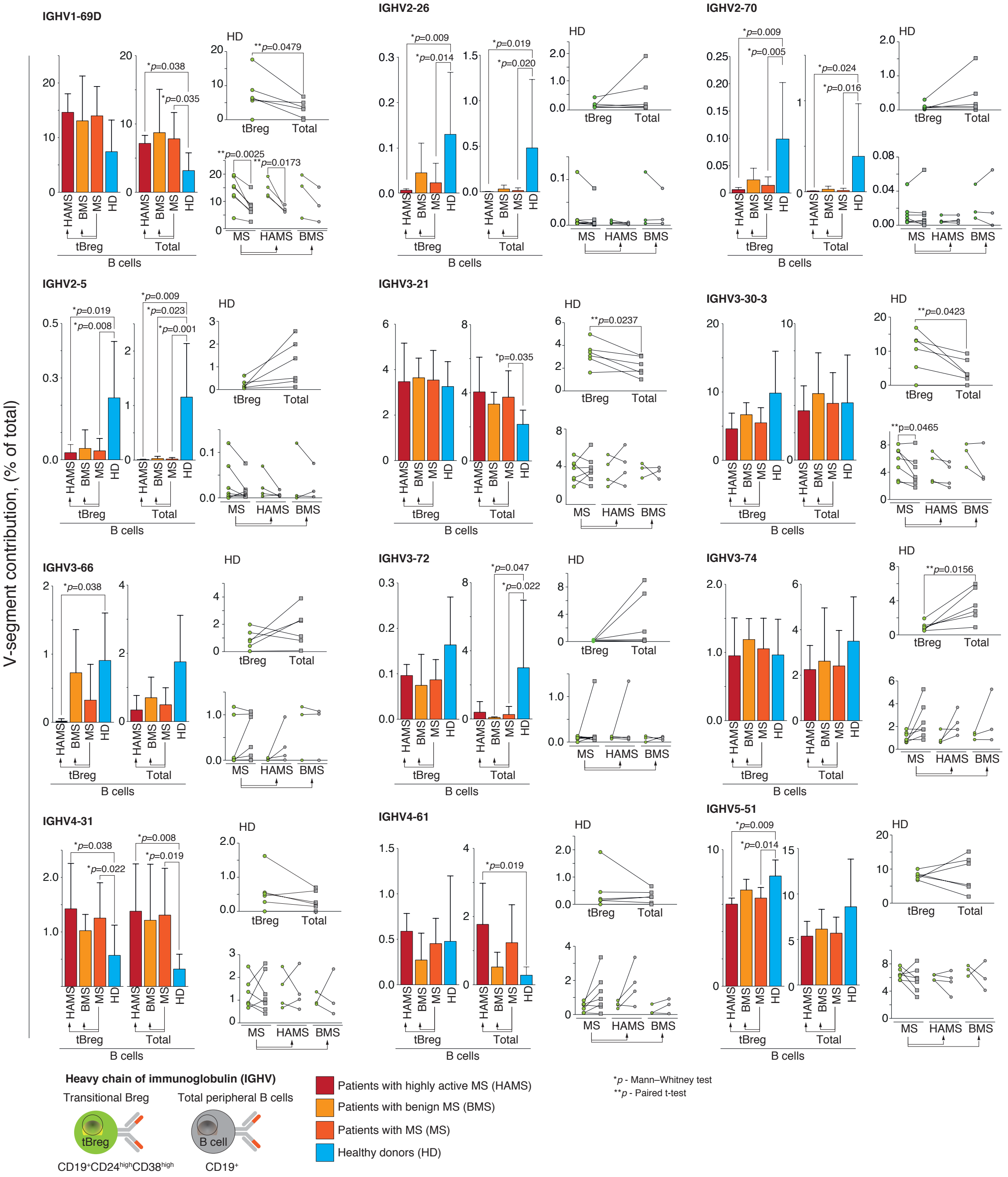


Fig. 1. Differential usage of germline Ig VH-gene segments in MS patients and healthy donors. The frequency of 49 functional VH genes in patients with MS, separately for BMS and HAMS, was analyzed. The frequency of VH germline genes in healthy donors (HD) was analyzed as a control. The distribution of the germline gene repertoire was compared between the total pool of peripheral blood B cells (CD19+) and tBregs with the CD19$^+$CD24^{high}CD38^{high} phenotype. Histograms show the comparison of patients with different types of MS courses to healthy donors, where the average value (mean ± SD) of the proportion of IgG sequences related to the indicated germline is provided for each group. The individual values of the proportion of immunoglobulin germline genes for each patient are compared for the total pool of B cells (total, gray dots) and the subpopulation of transient regulatory B cells (tBregs, green dots) and shown to the right of the histogram. The data are provided only for the germline genes for which a statistically significant difference was shown in at least one analyzed parameter (comparison of different types of MS courses against healthy donors was performed using the Mann–Whitney test; comparison of the total pool of B cells with the tBreg subpopulation was performed using the paired t-test; only statistically significant *p*-values are shown)

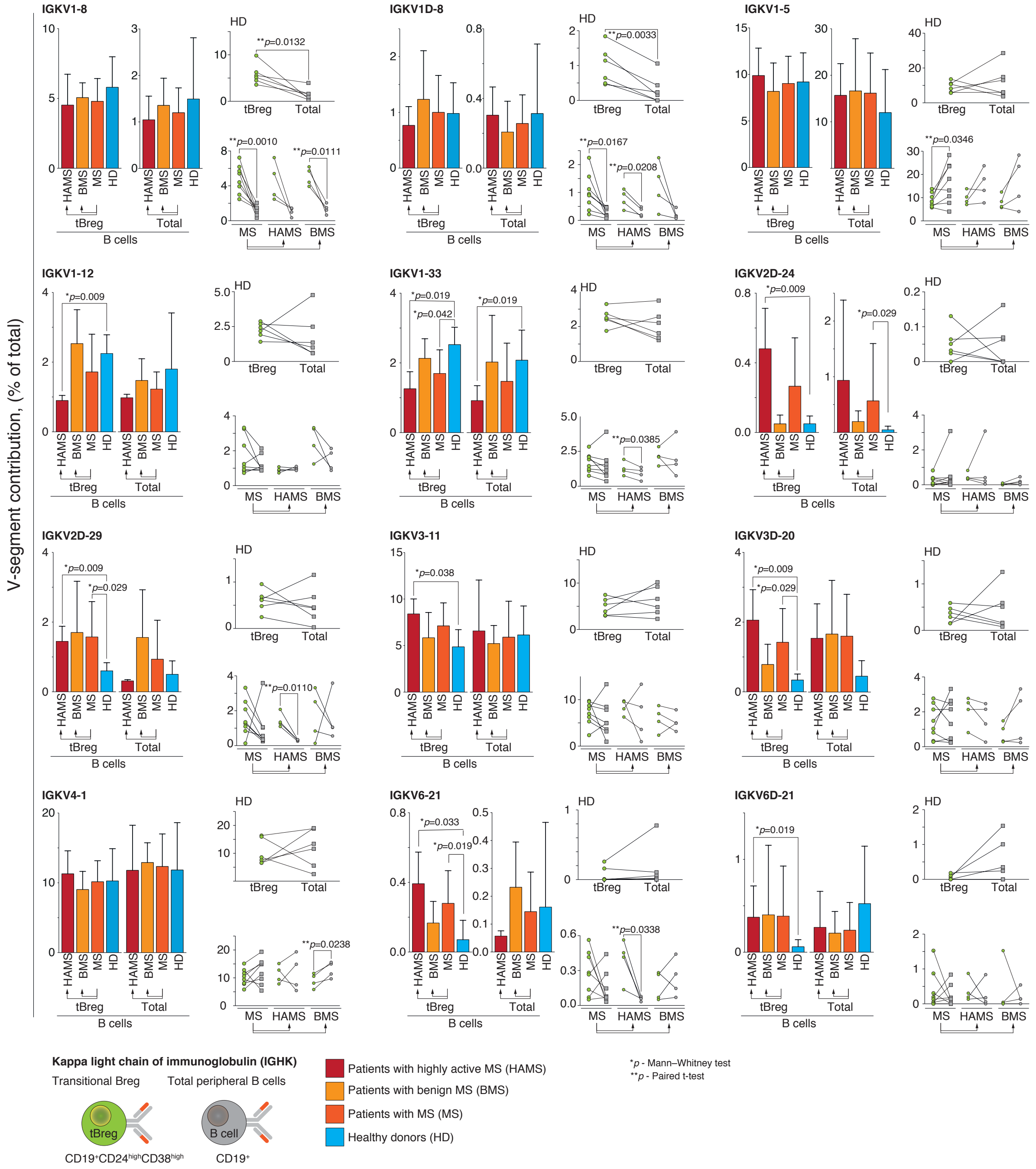


Fig. 2. Differential usage of germline Ig Vk-gene segments in MS patients and healthy donors. Frequency of 41 functional Vk genes in MS patients, separately for BMS and HAMS patients, was analyzed. The frequency of Vk germline genes in healthy donors (HD) was analyzed as a control. The distribution of the germline gene repertoire was compared between the total pool of peripheral blood B cells (CD19+) and tBregs with the CD19$^+$CD24^{high}CD38^{high} phenotype. Histograms show the comparison of patients with different types of MS course and healthy donors, where the average value (mean ± SD) of the proportion of IgG sequences related to the indicated germline is provided for each group of patients. Individual values of the proportion of immunoglobulin germline genes for each patient are compared for the total pool of B cells (total, gray dots) and the subpopulation of transient regulatory B cells (tBregs, green dots) and shown to the right of the histogram. The data are provided only for germline genes, for which a statistically significant difference was shown in at least one analyzed parameter (comparison of patients with different types of MS course against healthy donors was performed using the Mann–Whitney test; comparison of the total pool of B cells with the subpopulation of tBregs was performed using the paired t-test; only statistically significant *p*-values are shown)

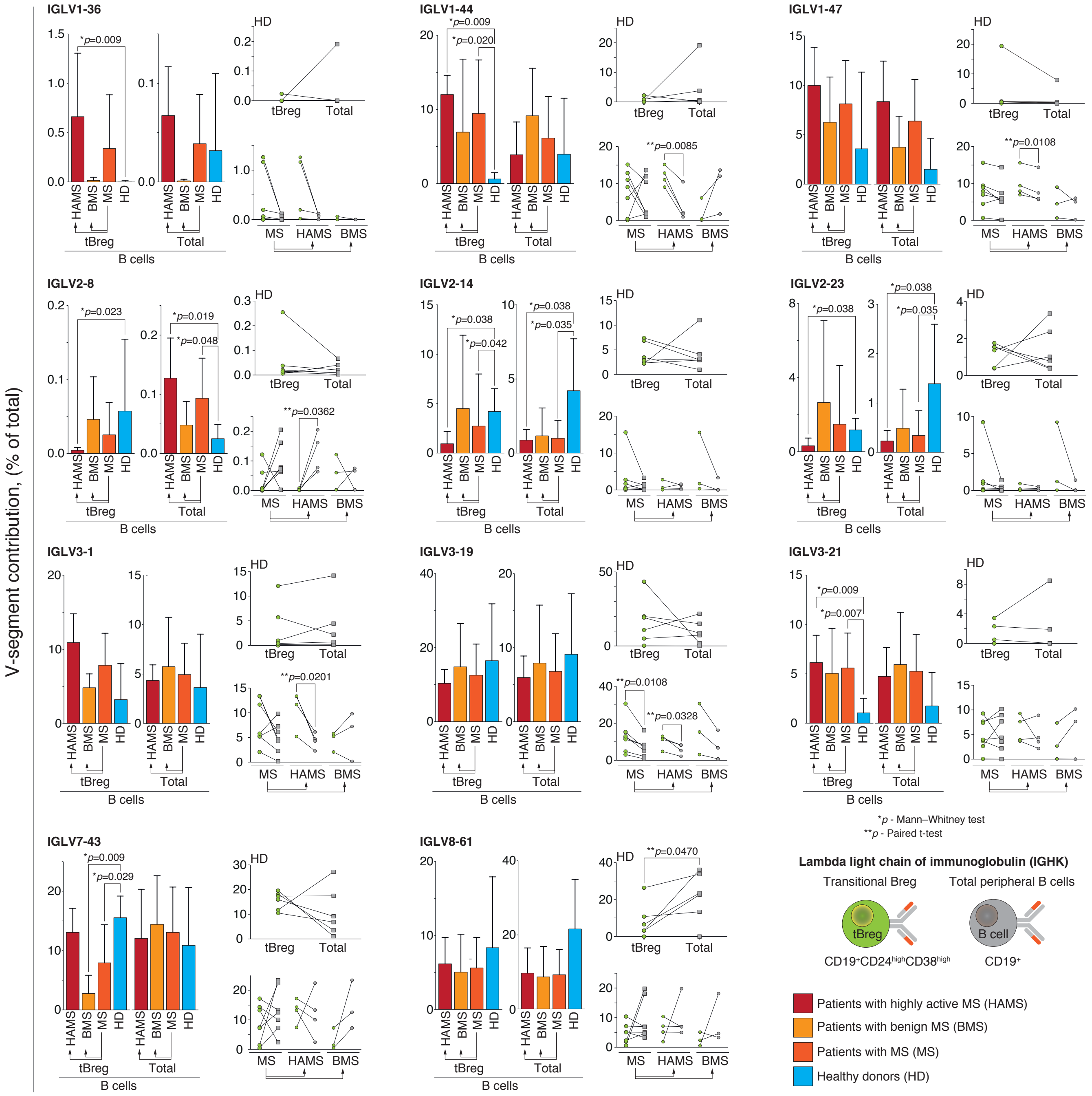


Fig. 3. Differential usage of germline Ig Vλ-gene segments in MS patients and healthy donors. Frequency of 26 functional Vλ genes in MS patients, separately for BMS and HAMS patients, was analyzed. The frequency of Vλ germline genes in healthy donors (HD) was analyzed as a control. The distribution of the germline gene repertoire was compared between the total pool of peripheral blood B cells (CD19+) and tBregs with the CD19$^+$CD24^{high}CD38^{high} phenotype. Histograms show a comparison of patients with different types of MS courses and healthy donors, where the average value (mean ± SD) of the proportion of IgG sequences related to the indicated germline is provided for each group. Individual values of the proportion of immunoglobulin germline genes for each patient are compared for the total pool of B cells (total, gray dots) and the subpopulation of transient regulatory B cells (tBregs, green dots) and shown to the right of the histogram. The data are provided only for the germline genes for which a statistically significant difference was shown in at least one analyzed parameter (comparison of patients with different types of MS courses against healthy donors was performed using the Mann–Whitney test; comparison of the total pool of B cells with the subpopulation of tBregs was performed using the paired t-test; only statistically significant *p*-values are shown)

germline genes in the tBreg population does not significantly differ between BMS patients and healthy donors.

Differences in the distribution of immunoglobulin germline genes in tBregs during MS development are also observed in the case of the lambda light chain isotype (*Fig. 3*). The IGLV1-36 germline is almost never observed in healthy donors and BMS patients, but its frequency significantly rises to 0.5% in HAMS patients. The frequency of IGLV1-44 and IGLV3-21 germline genes is increased in patients with any type of MS course; however, a statistically significant difference is observed only between HAMS patients and healthy donors. The distribution of IGLV2-8, IGLV2-14, and IGLV2-23 germline genes does not differ between BMS patients and healthy donors, but their frequency significantly decreases with the development of HAMS. Interestingly, representation of the IGLV7-43 germline gene, conversely, is approximately the same in HAMS patients and healthy donors, but significantly decreases in BMS patients.

CONCLUSIONS

Immunological studies carried out in the 21st century have confirmed the crucial role of Bregs in maintaining immunotolerance, as well as controlling and reducing the inflammatory response. There are still many questions regarding the exact mechanism of its regulation, but it is obvious that a violation of the number and function of Breg cells leads to the development of various immunological pathologies, among which MS is particularly prominent. A detailed elucidation of inflammatory regulation by B cells will allow us not only to determine the etiology of autoimmune pathologies, but also may contribute to the development of Breg-based therapy in the near future. Immunoglobulins play an important role in the immune response by being exposed as antigen-specific receptors on the B-cell surface, as well as secreted antibodies. The recent progress achieved in the NGS analysis makes it possible to identify immunoglobulin repertoires with an unprecedented high level of detailing [32]. Therefore, it is extremely important to study the structure and functions of immunoglobulins, their specificity, and epigenetic status to understand the fundamental principles of MS onset and progression. Over the recent years, more and more patterns and stereotyped antibody responses have been discovered, when different individuals produce immunoglobulins recognizing certain antigenic epitopes using the same IgV genes [32–34]. In other words, certain immunoglobulin germlines exhibit tropism to certain antigens. Accordingly, variations in the usage of some immunoglobulin germline genes can be associated with a different antibody's ability to generate an effective immune response, which may manifest itself as predisposition to various diseases, including autoimmune ones. It is likely that the differences in the germline gene frequency in each individual can be the result of an antiviral or autoimmune response. Before the onset of antigen-dependent B cell differentiation mediated by somatic hypermutation of immunoglobulin sequences, the diversity of immature B cells, including tBregs, is almost entirely determined by the configuration of the body's germline genes (V(D)J recombination). Therefore, a detailed study of the immunoglobulin repertoire of immature B cells in patients with various autoimmune diseases, including MS, will help determine which rearrangements in the immunoglobulin germline genes can lead to functional disorders of the immune system.

The present study revealed that the distribution of immunoglobulin germline genes in the tBreg population in MS patients differs from that in a healthy person. Particularly significant differences are observed for the IGLV1-44 and IGHV2-5 germlines. The IGLV1-44 lambda chain germline is almost absent in the tBreg subpopulation of a healthy person but is found in MS patients. The IGHV4-31 germline is more frequent during MS development, both in the total pool of B cells and in the tBreg subpopulation. The opposite situation is observed for the IGHV2-26, IGHV2-5, IGHV2-70 germline genes of the heavy chain: these germlines, although being identified at small amounts in almost every healthy individual, disappear in MS patients. Moreover, in the case of a severe form of the disease, the difference from the normal value becomes larger. We have also previously found more significant differences in the number of peripheral tBregs and their maturation level in HAMS patients compared with BMS patients and healthy donors [20]. Therefore, this study has shown that a more significant variation in the tBreg $CD19^{+}CD38^{high}CD24^{high}$ subpopulation repertoire is associated with a more aggressive MS course. In general, a similar situation is observed for all the germlines but IGLV1-44: if the immunoglobulin germline frequency in the total pool of circulating B cells changes during MS development, a similar picture is also true for tBregs. Therefore, during the development of the autoimmune MS pathology, disruptions in the distribution of the immunoglobulin germline genes can be genetically predetermined and occur already at an early stage of B-cell maturation. To confirm this hypothesis, the size of the analyzed patients' cohorts needs to be increased and the differences in the structure and specificity of the B-cell receptors of other Breg subpopulations need to be studied. ●

This study was supported by the Russian Science Foundation (grant No.18-74-10079 "Self-assembled genetically encoded nanocages as a tool for the treatment of multiple sclerosis"), the Russian Foundation for Basic Research grant No. 17-00-00229 (obtaining peripheral blood of patients with multiple sclerosis and healthy donors) and project No. 075-15-2021-1033 (13.2251.21.0111) in terms of large-scale sequencing.

Cell sorting was carried out using the equipment provided by the Center for Collective Use of the Institute of Bioorganic Chemistry and supported by the Ministry of Science and Higher Education of the Russian Federation (project No. 075-15-2020-807).

REFERENCES

1. Thompson A.J., Baranzini S.E., Geurts J., Hemmer B., Ciccarelli O. // Lancet. 2018. V. 391. P. 1622–1636. https://doi.org/10.1016/S0140-6736(18)30481-1.
2. Belogurov A., Kuzina E., Kudriaeva A., Kononikhin A., Kovalchuk S., Surina Y., Smirnov I., Lomakin Y., Bacheva A. Stepanov A., et al. // FASEB J. 2015. V. 29. P. 1901–1913. https://doi.org/10.1096/fj.14-259333.
3. Belogurov A.A., Stepanov A.V., Smirnov I.V., Melamed D., Bacon A., Mamedov A.E., Boitsov V.M., Sashchenko L.P., Ponomarenko N.A., Sharanova S.N., et al. // FASEB J. 2013. V. 27. P. 222–231. https://doi.org/10.1096/fj.12-213975.
4. Belogurov A., Kudriaeva A., Kuzina E., Smirnov I., Bobik T., Ponomarenko N., Kravtsova-Ivantsiv Y., Ciechanover A., Gabibov A. // J. Biol. Chem. 2014. V. 289. P. 17758–17766. https://doi.org/10.1074/JBC.M113.544247.
5. Baecher-Allan C., Kaskow B.J., Weiner H.L. // Neuron. 2018. V. 97. P. 742–768. https://doi.org/10.1016/j.neuron.2018.01.021.
6. Ziganshin R.H., Ivanova O.M., Lomakin Y.A., Belogurov A.A., Kovalchuk S.I., Azarkin I.V., Arapidi G.P., Anikanov N.A., Shender V.O., Piradov M.A., et al. // Mol. Cell. Proteomics. 2016. V. 15. P. 2366–2378. https://doi.org/10.1074/mcp.M115.056036.
7. Ramasamy R., Mohammed F., Meier U.C. // Immunol. Lett. 2020. V. 217. P. 15–24. https://doi.org/10.1016/j.imlet.2019.10.017.
8. Wekerle H., Hohlfeld R. // N. Engl. J. Med. 2003. V. 349. P. 185–186. https://doi.org/10.1056/NEJMcibr035136.
9. Lomakin Y., Arapidi G.P., Chernov A., Ziganshin R., Tcyganov E., Lyadova I., Butenko I.O., Osetrova M., Ponomarenko N., Telegin G., et al. // Front. Immunol. 2017. V. 8. https://doi.org/10.3389/fimmu.2017.00777.
10. Lanz T.V., Brewer R.C., Ho P.P., Moon J.-S., Jude K.M., Fernandez D., Fernandes R.A., Gomez A.M., Nadj G.S., Bartley C.M., et al. // Nature. 2022. V. 603. P. 321–327. https://doi.org/10.1038/s41586-022-04432-7.
11. Bjornevik K., Cortese M., Healy B.C., Kuhle J., Mina M.J., Leng Y., Elledge S.J., Niebuhr D.W., Scher A.I., Munger K.L., et al. // Science. 2022. V. 375. P. 296–301. https://doi.org/10.1126/science.abj8222.
12. Gabibov A.G., Ponomarenko N.A., Tretyak E.B., Paltsev M.A., Suchkov S.V. // Autoimmun. Rev. 2006. V. 5. P. 324–330. https://doi.org/10.1016/j.autrev.2006.01.004.
13. Ponomarenko N.A., Durova O.M., Vorobiev I.I., Belogurov A.A., Kurkova I.N., Petrenko A.G., Telegin G.B., Suchkov S.V., Kiselev S.L., Lagarkova M.A., et al. // Proc. Natl. Acad. Sci. USA. 2006. V. 103. P. 281–286. https://doi.org/10.1073/pnas.0509849103.
14. Lomakin Y., Kudriaeva A., Kostin N., Terekhov S., Kaminskaya A., Chernov A., Zakharova M., Ivanova M., Simaniv T., Telegin G., et al. // Sci. Rep. 2018. V. 8. P. 12679. https://doi.org/10.1038/s41598-018-30938-0.
15. Gabibov A.G., Belogurov A.A., Lomakin Y.A., Zakharova M.Y., Avakyan M.E., Dubrovskaya V.V., Smirnov I.V., Ivanov A.S., Molnar A.A., Gurtsevitch V.E., et al. // FASEB J. 2011. V. 25. P. 4211–4221. https://doi.org/10.1096/fj.11-190769.
16. Lomakin Y.A., Zakharova M.Y., Stepanov A.V., Dronina M.A., Smirnov I.V., Bobik T.V., Pyrkov A.Y., Tikunova N.V., Sharanova S.N., Boitsov V.M., et al. // Mol. Immunol. 2014. V. 62. P. 305–314. https://doi.org/10.1016/j.molimm.2014.01.013.
17. Sokolov A.V., Shmidt A.A., Lomakin Y.A. // Acta Naturae. 2018. V. 10. P. 11–22.
18. Ran Z., Yue-Bei L., Qiu-Ming Z., Huan Y. // Front. Immunol. 2020. V. 11. https://doi.org/10.3389/fimmu.2020.01884.
19. Matsumoto M., Fujii Y., Baba A., Hikida M., Kurosaki T., Baba Y. // Immunity. 2011. V. 34. P. 703–714. https://doi.org/10.1016/j.immuni.2011.03.016.
20. Lomakin Y.A., Zvyagin I.V., Ovchinnikova L.A., Kabilov M.R., Staroverov D.B., Mikelov A., Tupikin A.E., Zakharova M.Y., Bykova N.A., Mukhina V.S., et al. // Front. Immunol. 2022. V. 13. P. 3678. https://doi.org/10.3389/fimmu.2022.803229.
21. Díaz C., Zarco L.A., Rivera D.M. // Mult. Scler. Relat. Disord. 2019. V. 30. P. 215–224. https://doi.org/10.1016/j.msard.2019.01.039.
22. Schaefer L.M., Poettgen J., Fischer A., Gold S., Stellmann J.P., Heesen C. // Brain Behav. 2019. V. 9. P. e01259. https://doi.org/10.1002/brb3.1259.
23. Sand I.K., Krieger S., Farrell C., Miller A.E. // Mult. Scler. J. 2014. V. 20. P. 1654–1657. https://doi.org/10.1177/1352458514521517.
24. Cheng J., Torkamani A., Grover R.K., Jones T.M., Ruiz D.I., Schork N.J., Quigley M.M., Hall F.W., Salomon D.R., Lerner R.A. // Proc. Natl. Acad. Sci. USA. 2011. V. 108. P. 560–565. https://doi.org/10.1073/pnas.1101148108.
25. Bolotin D.A., Poslavsky S., Mitrophanov I., Shugay M., Mamedov I.Z., Putintseva E.V., Chudakov D.M. // Nat. Meth. 2015. V. 12. P. 380–381. https://doi.org/10.1038/nmeth.3364.
26. Choi J.K., Yu C.R., Bing S.J., Jittayasothorn Y., Mattapallil M.J., Kang M., Park S.B., Lee H.S., Dong L., Shi G., et al. // Proc. Natl. Acad. Sci. USA. 2021. V. 118. https://doi.org/10.1073/PNAS.2109548118.
27. Radomir L., Kramer M.P., Perpinial M., Schottlender N., Rabani S., David K., Wiener A., Lewinsky H., Becker-

Herman S., Aharoni R., et al // Nat. Commun. 2021. V. 12. https://doi.org/10.1038/S41467-021-22230-Z.

28. Blair P.A., Noreña L.Y., Flores-Borja F., Rawlings D.J., Isenberg D.A., Ehrenstein M.R., Mauri C. // Immunity. 2010. V. 32. P. 129–140. https://doi.org/10.1016/j.immuni.2009.11.009.

29. Zhu H.-Q., Xu R.-C., Chen Y.-Y., Yuan H.-J., Cao H., Zhao X.-Q., et al. // Br. J. Dermatol. 2015. V. 172. P. 101–110. https://doi.org/10.1111/bjd.13192.

30. von Büdingen H.C., Kuo T.C., Sirota M., van Belle C.J., Apeltsin L., Glanville J., Cree B.A., Gourraud P.A., Schwartzburg A., Huerta G., et al. // J. Clin. Invest. 2012. V. 122. P. 4533–4543. https://doi.org/10.1172/JCI63842.

31. Owens G.P., Winges K.M., Ritchie A.M., Edwards S., Burgoon M.P., Lehnhoff L., Nielsen K., Corboy J., Gilden D.H., Bennett J.L. // J. Immunol. 2007. V. 179. P. 6343–6351. https://doi.org/10.4049/jimmunol.179.9.6343.

32. Mikocziova I., Greiff V., Sollid L.M. // Genes Immun. 2021. V. 22. P. 205–217. https://doi.org/10.1038/s41435-021-00145-5.

33. Henry Dunand C.J., Wilson P.C. // Philos. Trans. R Soc. Lond. B Biol. Sci. 2015. V. 370. https://doi.org/10.1098/RSTB.2014.0238.

34. Wang Y., Yuan M., Lv H., Peng J., Wilson I.A., Wu N.C. // Immunity. 2022. V. 55. P. 1105–1117.e4. https://doi.org/10.1016/j.immuni.2022.03.019.

A Low-Molecular-Weight BDNF Mimetic, Dipeptide GSB-214, Prevents Memory Impairment in Rat Models of Alzheimer's Disease

P. Yu. Povarnina[1*], A. A. Volkova[1,2], O. N. Vorontsova[1], A. A. Kamensky[2], T. A. Gudasheva[1], S. B. Seredenin[1]

[1]Research Zakusov Institute of Pharmacology, Moscow, 125315 Russia

[2]Lomonosov Moscow State University, Faculty of Biology, Moscow, 119991 Russia

[*]E-mail: povarnina@gmail.com

Received June 16, 2022; in final form, September 30, 2022

DOI: 10.32607/actanaturae.11755

ABSTRACT Brain-derived neurotrophic factor (BDNF) is known to be involved in the pathogenesis of Alzheimer's disease (AD). However, the pharmacological use of full-length neurotrophin is limited, because of its macromolecular protein nature. A dimeric dipeptide mimetic of the BDNF loop 1, bis-(N-monosuccinyl-*L*-methionyl-*L*-serine) heptamethylene diamide (GSB-214), was designed at the Zakusov Research Institute of Pharmacology. GSB-214 activates TrkB, PI3K/AKT, and PLC-γ1 *in vitro*. GSB-214 exhibited a neuroprotective activity during middle cerebral artery occlusion in rats when administered intraperitoneally (i.p.) at a dose of 0.1 mg/kg and improved memory in the novel object recognition test (0.1 and 1.0 mg/kg, i.p.). In the present study, we investigated the effects of GSB-214 on memory in the scopolamine- and steptozotocin-induced AD models, with reference to activation of TrkB receptors. AD was modeled in rats using a chronic i.p. scopolamine injection or a single streptozotocin injection into the cerebral ventricles. GSB-214 was administered within 10 days after the exposure to scopolamine at doses of 0.05, 0.1, and 1 mg/kg (i.p.) or within 14 days after the exposure to streptozotocin at a dose of 0.1 mg/kg (i.p.). The effect of the dipeptide was evaluated in the novel object recognition test; K252A, a selective inhibitor of tyrosine kinase receptors, was used to reveal a dependence between the mnemotropic action and Trk receptors. GSB-214 at doses of 0.05 and 0.1 mg/kg statistically significantly prevented scopolamine-induced long-term memory impairment, while not affecting short-term memory. In the streptozotocin-induced model, GSB-214 completely eliminated the impairment of short-term memory. No mnemotropic effect of GSB-214 was registered when Trk receptors were inhibited by K252A.

KEYWORDS brain-derived neurotrophic factor, dimeric dipeptide mimetic, Alzheimer's disease, scopolamine, streptozotocin, memory.

ABBREVIATIONS BDNF – brain-derived neurotrophic factor; SC – scopolamine; STZ – streptozotocin; AD – Alzheimer's disease.

INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia, accounting for 60–80% of all dementia cases, while no effective pathogenetic therapy exists today for this disease [1].

Over the past two decades, regulation of the activity of neurotrophin receptors, and the brain-derived neurotrophic factor (BDNF) in particular, has been viewed as a new strategy for treating neurodegenerative diseases. BDNF maintains neuronal viability and synaptic plasticity, playing an important role in the processes of learning and memory. Data indicative of BDNF involvement in the pathogenesis of AD have been published [2–4]. Reduced BDNF expression is already observed at the early stage of the disease and correlates with an accumulation of β-amyloid and the hyperphosphorylated tau protein [5]. The favorable effects of exogenous BDNF have been demonstrated in various AD models. BDNF ensures neuronal protection under conditions of β-amyloid toxicity both *in vitro* and *in vivo* [6]. Insertion of the *BDNF* gene within a lentiviral vector into J20 transgenic mice (carrying mutations in the gene encoding the amyloid precursor protein) prevented the death of the cells of the entorhinal cortex and improved cognitive functions [7]. It has been shown using another genetic model of AD (P301L mice carrying the mutant tau protein gene) that stable human *BDNF* gene expression restored the BDNF level, thus preventing neuronal and synaptic degeneration in the hippocampus, as well as cognitive disorders [8]. However, the gene therapy has such shortcomings as invasiveness, high cost, and the risk of adverse effects related to the pleiotropic effect of BDNF.

The clinical use of BDNF is impeded by its poor penetration through the blood–brain barrier and rapid degradation [9]. Low-molecular-weight BDNF mimetics with improved pharmacokinetic properties are currently being developed [10, 11]. Activity of the low-molecular-weight BDNF mimetic 7,8-dihydroxyflavone, a TrkB receptor agonist, was determined using AD models [12–14].

A dimeric dipeptide mimetic of the BDNF loop 1, GSB-214 (bis-(N-monosuccinyl-*L*-methionyl-*L*-serine) heptamethylene diamide), was designed and synthesized at the Zakusov Research Institute of Pharmacology based on the hypothesis that the most exposed domains of the loop-like neurotrophin structures (most frequently, the central domains of their β turns) exhibit pharmacophoric properties [15] [RU Patent 2410392, 2011; US Patent 9683014 B2, 2017; CN Patent 102365294 B, 2016; EU Patent 2397488, 2019; IN Patent 296506, 2018] (*Fig. 1*).

Earlier, Western blotting showed that incubation of HT-22 mouse hippocampal cells in the presence of GSB-214 for 5–180 min results in the activation of TrkB receptors and the conjugated PI3K/Akt and PLC-γ1 signaling pathways, but not the MAPK/ERK signaling pathway [10]. It has been shown using HT-22 cells that GSB-214 at micro-nanomolar concentrations exhibits neuroprotective activity under oxidative stress [15].

The dipeptide GSB-214 (administered i.p. at doses of 0.1–0.5 mg/kg) exhibited *in vivo* neuroprotective activity in a rat model of transient middle cerebral artery occlusion [16] and antidiabetic activity in a streptozotocin-induced model of diabetes in mice [17]. Taking into account the findings regarding the similarity of the pathogenesis of diabetes and AD [18], the antidiabetic properties of GSB-214, along with the neuroprotective properties, indicate that there is promise in studying the effects of the dipeptide in AD models.

The objective of our work was to investigate the effect of GSB-214 on memory in the scopolamine- and streptozotocin-induced models of AD, as well as evaluate its mnemotropic activity as a function of the activation of Trk receptors.

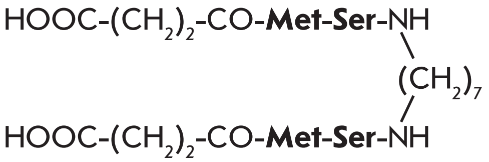


Fig. 1. The dimeric dipeptide mimetic of the BDNF loop 1 GSB-214

EXPERIMENTAL

Materials

The dipeptide GSB-214 was synthesized at the Medicinal Chemistry Department of the Zakusov Research Institute of Pharmacology according to the procedure described earlier [14]; 96% chromatographic purity (HPLC), $[\alpha]^{25}_{D}$ = +9.0° (0.4 in DMF), T_{melt} = 162–163°C. Scopolamine (Acros Organics, USA), streptozotocin, and K252A (Sigma Aldrich, USA) were used.

Animals

The experiments were conducted using male Wistar rats (weight, 230–260 g) procured from the Andreevka Branch of the Research Center for Biomedical Technologies, the Federal Medical-Biological Agency (FMBA). The animals were kept in a vivarium with *ad libitum* feeding and access to water and natural light–dark cycle. The behavioral experiments were carried out at a time interval between 10 a.m. and 2 p.m. (local time). The animal experiments were carried out in compliance with international regulations (Directive 2010/63/EU of the European Parliament and of the Council of the European Union of September 22, 2010, on the protection of animals used for scientific purposes). The experiments were approved by the Biomedical Ethics Committee of the Zakusov Research Institute of Pharmacology (Protocol No. 3 dated February 18, 2021).

Scopolamine-induced model of AD

The rats were randomly assigned to the following groups: Control (n = 9), Scopolamine (SC) (n = 10), SC + GSB-214 (0.05 mg/kg) (n = 10), SC + GSB-214 (0.1 mg/kg) (n = 9), and SC + GSB-214 (1.0 mg/kg) (n = 10). Scopolamine in normal saline was injected i.p. to rats at a dose of 2 mg/kg during 20 days. GSB-214 in distilled water was injected i.p. at doses of 0.05, 0.1, and 1.0 mg/kg during 10 days after exposure to scopolamine. The rats in the Control group received equiva-

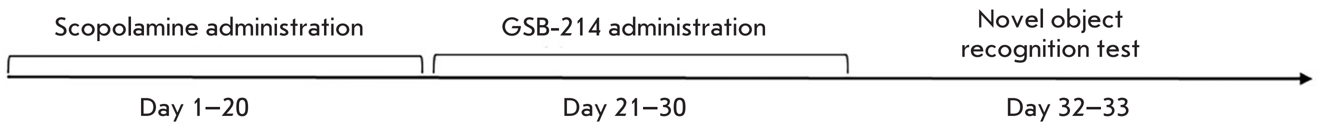


Fig. 2. The scheme of the experiment on the mnemotropic effects of GSB-214 in the scopolamine-induced AD model

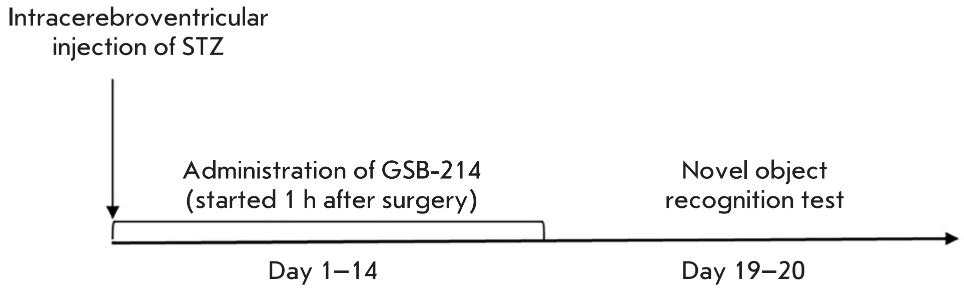

Intracerebroventricular injection of STZ
Administration of GSB-214 (started 1 h after surgery)
Novel object recognition test
Day 1–14
Day 19–20

Fig. 3. The scheme of the experiment on the mnemotropic effects of GSB-214 in the streptozotocin-induced AD model. STZ – Streptozotocin

lent volumes of normal saline, instead of scopolamine, and distilled water, instead of GSB-214, according to the same scheme. The rats in the SC group received scopolamine and distilled water.

The novel object recognition test was carried out on days 32–33.

The scheme of the experiment is shown in *Fig. 2*.

Streptozotocin-induced model of AD

The rats were randomly assigned to the following groups: Control ($n = 10$), Streptozotocin (STZ) ($n = 7$), and STZ + GSB-214 (0.1 mg/kg) ($n = 8$). STZ in citrate buffer was stereotactically injected into the cerebral ventricles at a dose of 3 mg/kg (AP = −1.0; L = 1.5; depth, 3.5). The injection volume was 3 μL per ventricle; the injection rate was 1 μL/min. One hour after the exposure, the rats received an i.p. injection of GSB-214 (0.1 mg/kg) and, then, received injections once daily during 13 days. The rats in the Control group were injected with equivalent volumes of citrate buffer, instead of STZ, and distilled water, instead of GSB-214, according to the same scheme. The rats in the STZ group received STZ and distilled water.

The novel object recognition test was carried out on days 19–20. The scheme of the experiment is shown in *Fig. 3*.

The novel object recognition test

This test is based on the natural rodents' instinct to investigate novel objects [19]. It is widely used for assessing both short-term and long-term memory [20].

The test was conducted in T4 cages identical to the home cages where the animals had been housed throughout the study. A rat was first placed into an empty cage with the floor covered with sawdust for 4 min to adapt.

The familiarization phase. Two identical objects not familiar to the rat were placed in the two nearest corners of the cage. The time spent exploring the objects was recorded during 4 min. The rat was then returned to its home cage.

Test. A new pair of objects was placed in the same corners of the cage; one object was identical to those presented to the rats during the familiarization phase, while the other was unfamiliar. The time spent exploring the familiar and novel objects was recorded during 4 min. The test was carried out 1 h (test 1) and 24 h (test 2) after the familiarization phase to record the short-term and long-term memory, respectively. Different unfamiliar objects were used in test 1 and test 2. Exploration was defined as sniffing, with the distance between the animal's snout and the object being ⩽ 2 cm.

The discrimination index was used as the memory criterion [21]; it was calculated using the formula: $DI = (T_{\text{novel}} - T_{\text{fam}})/(T_{\text{novel}} + T_{\text{fam}})$, where T_{novel} was the time spent exploring a novel object and T_{fam} was the time spent exploring a familiar object. The K_D values > 0 meant that the animal remembered the object presented to it at the familiarization phase.

Pharmacological inhibitory analysis

The rats were randomly assigned to the following groups: Control (distilled water and 1% DMSO in normal saline, $n = 12$), GSB-214 0.1 mg/kg (GSB-214 and 1% DMSO, $n = 13$), GSB-214 0.1 mg/kg + K252A 100 μg/kg ($n = 12$), and K252A 100 μg/kg (distilled water and K252A, $n = 13$). GSB-214 at a dose of 0.1 mg/kg or an equivalent amount of distilled water was administered i.p. 20 min after the i.p. injection of K252A (100 μg/kg) in 1% DMSO or 1% DMSO. The novel object recognition test was started after 24 h. The dose of GSB-214 was chosen based on earlier experiments [22].

Statistical analysis

Statistical analysis of the experimental data was performed using the GraphPad Prism 8.0 software (GraphPad Software, USA). The statistical significance of differences in the discrimination index was assessed using one-way ANOVA, followed by pairwise intergroup comparisons using the Dunnett's test or two-factor ANOVA followed by pairwise intergroup comparisons using the Tukey's test.

The data were presented as the mean ± standard error of the mean. Differences were considered statistically significant at $p < 0.05$.

RESULTS

The dipeptide GSB-214 prevents long-term memory impairment in the scopolamine-induced model of AD

Compared to the control group, chronic administration of scopolamine significantly reduced the discrimination index in both test 1 (1 h after becoming familiar with the objects, $p = 0.0212$) and test 2 (24 h after becoming familiar with the objects, $p = 0.0077$), thus indicating that short-term and long-term memory, respectively, was impaired (*Table 1*). Chronic administration of GSB-214 at doses of 0.05 and 0.1 mg/kg prevented long-term memory impairment ($p = 0.0177$ and 0.0304 vs. SC group, respectively), although it had no effect on short-term memory. No activity was observed for the dipeptide GSB-214 when administered at a dose of 1.0 mg/kg (*Table 1*).

Hence, GSB-214 i.p. administered at doses of 0.05 and 0.1 mg/kg for 10 days proved effective against long-term memory impairment in the scopolamine-induced model of AD.

The dipeptide GSB-214 prevents short-term memory impairment in a streptozotocin-induced model of AD

In the streptozotocin-induced model of AD, we uncovered significant memory impairment in the rats in the STZ group 1 h after becoming familiar with the objects ($p = 0.0045$), but not after 24 h (*Table 2*). Therefore, in this experimentally induced model of AD, rats experienced short-term, rather than long-term, memory impairment, which is typical of the early stage of the disease [23]. GSB-214 at a dose of 0.1 mg/kg yielded a statistically significant correction of this impairment ($p = 0.0032$); the discrimination index in the group of animals receiving treatment was 4.8-fold higher compared to that in the STZ group (*Table 2*).

Hence, the dipeptide GSB-214 completely inhibited short-term memory impairment in the streptozotocin-induced model of AD.

The mnemotropic activity of GSB-214 depends on the activation of Trk receptors

In order to confirm the involvement of the activation of Trk receptors in the mnemotropic effects of GSB-214, we studied how K252A, an inhibitor of these receptors, influences the effects of GSB-214 in the novel object recognition test. *Table 3* shows that the dipeptide GSB-214 significantly improved long-term memory as the discrimination index in the test after 24 h in this case increased approximately 1.5-fold compared to that in the control group. This effect was completely eliminated by injecting a K252A inhibitor 20 min before the exposure to GSB-214. K252A per se did not affect the rats' memory. The studied compounds were found to exhibit no effect on the short-term memory of the rats (test 1) (*Table 3*).

DISCUSSION

Earlier, we had found that a single-dose BDNF dipeptide mimetic GSB-214 administered i.p. (0.1 and 1.0 mg/kg) had a favorable effect on the long-term memory of rats in the novel object recognition test [22].

In this study, we investigated the mnemotropic activity of GSB-214 in the same test in the scopolamine- and streptozotocin-induced models of AD.

The scopolamine-induced amnesia model is commonly used for evaluating potential therapeutic agents for treating AD [24–26]. Chronic exposure to scopolamine causes cholinergic deficit that is mainly induced by blockade of acetylcholine receptors and, therefore, cognitive impairment [25]. In our modification of the model [24], the impairment induced by chronic exposure to scopolamine and its subsequent discontinuation (see the scheme of the experiment in *Fig. 2*) is attributed to the activation of feedback mechanisms, which first increase the density and affinity of acetylcholine receptors and subsequently induce the cholinergic deficit due to accelerated binding of the "available" acetylcholine.

The model of AD induced by intracerebroventricular injection of streptozotocin is also com-

Table 1. The effects of GSB-214 in the scopolamine-induced model of amnesia in the novel object recognition test

Group	Number of animals per group	Discrimination index	
		Test 1 (1 h)	Test 2 (24 h)
Control	9	0.57 ± 0.05	0.53 ± 0.06
SC	10	**0.3 ± 0.06***	**0.23 ± 0.06****
SC+GSB-214 (0.05 mg/kg)	10	0.48 ± 0.07	**0.48 ± 0.04**$^{\#}$
SC+GSB-214 (0.1 mg/kg)	9	0.45 ± 0.07	0.47 ± 0.05$^{\#}$
SC+GSB-214 (1.0 mg/kg)	10	0.33 ± 0.06	0.44 ± 0.08

The data are presented as the mean ± standard error of the mean. $^{**}p < 0.01$, $^{*}p < 0.05$ compared to the Control group; $^{\#}p < 0.05$ compared to the SC group (one-way ANOVA, the Dunnett's test).

Table 2. The effects of GSB-214 on short-term memory in the novel object recognition test for the streptozotocin-induced model of AD

Group	Number of animals per group	Discrimination index	
		Test 1 (1 h)	Test 2 (24 h)
Control	10	0.46 ± 0.07	0.49 ± 0.05
STZ	7	**0.1 ± 0.08****	0.43 ± 0.07
STZ+GSB-214 (0.1 mg/kg)	8	**0.48 ± 0.07**$^{\#\#}$	0.48 ± 0.03

The data are presented as the mean ± standard error of the mean. $^{**}p < 0.01$ compared to the Control group; $^{\#\#}p < 0.01$ compared to the STZ group (one-way ANOVA, the Dunnett's test).

Table 3. The Trk receptor inhibitor completely eliminates the mnemotropic effect of GSB-214 on long-term memory

Group	Number of animals per group	Discrimination index	
		Test 1 (1 h)	Test 2 (24 h)
Control	12	0.53 ± 0.07	0.47 ± 0.06
GSB-214 (0.1 mg/kg)	13	0.5 ± 0.05	0.73 ± 0.03***
GSB-214 (0.1 mg/kg) + K252A	12	0.53 ± 0.06	**0.36 ± 0.03**$^{\#\#\#\#}$
K252A	13	0.54 ± 0.06	**0.43 ± 0.05**

The data are presented as the mean ± standard error of the mean. $^{***}p < 0.001$ compared to the Control group; $^{\#\#\#\#}p < 0.0001$ compared to the GSB-214 group (two-way ANOVA, the Tukey's test).

monly used, has been validated, and studied well [27, 28]. Streptozotocin, a diabetogenic toxin, enters cells by binding to glucose transporter 2, because it is structurally similar to a sucrose molecule [28]. Intracerebral administration of streptozotocin induces insulin resistance and impairs brain glucose metabolism [29]. It causes neuropathological symptoms typical of AD, such as accumulation of β-amyloid and hyperphosphorylated tau protein, oxidative stress, as well as neuronal and synaptic death [30–33]. Like the scopolamine-induced model of AD, the streptozotocin-induced model is associated with memory disorders [31, 33].

We have revealed short-term and long-term memory impairment in the scopolamine-induced model of AD, which is consistent with the published data [26, 34]. The dipeptide GSB-214 eliminated only long-term memory impairment, while having no effect on short-term memory. This finding agrees with our earlier data obtained under physiological conditions in the novel object recognition test [22]. We assume that the revealed effect of GSB-214 can be attributed to the activation of the PI3K/Akt post-receptor signaling pathway, which was demonstrated earlier in *in vitro* experiments [10]. Serine/threonine protein kinase mTOR, one of the major protein synthesis regulators, is a component of the PI3K/Akt pathway [35]; it is viewed as the key factor in memory consolidation and, therefore, long-term memory formation [36]. It was found, using the novel object recognition test, that mTOR inhibition impairs long-term memory, but not short-term memory , in rats [37]. A hypothesis can be put forward that the effects of GSB-214 in the scopolamine-induced model of AD are related to the improvement of memory consolidation via the activation of the TrkB/PI3K/Akt/mTOR signaling pathway. We have demonstrated by pharmacological inhibitory analysis that the mnemotropic activity of GSB-214 is caused by an activation of the Trk neurotrophin receptors with which the PI3K/Akt/mTOR signaling pathway is associated.

In the streptozotocin-induced model, we observed only short-term memory impairment, which can be indicative of relatively mild neurodegenerative changes being characteristic of early AD [38]. GSB-214 eliminated this impairment. Since no effect of GSB-214 on short-term memory under physiological conditions was observed previously [22], it is fair to assume that memory was recovered due to the increase in neuronal viability under the exposure to streptozotocin-induced toxicity. The neuroprotective effects of GSB-214 were revealed earlier in *in vitro* experiments [15], as well as in a rat model of ischemic stroke induced by transient middle cerebral artery occlusion [16]. These effects, like the mnemotropic ones, are presumably associated with the activation of the PI3K/Akt signaling pathway. This pathway is known to mediate neuroprotection by inhibiting pro-apoptotic proteins and increasing the expression of anti-apoptotic proteins [39]. PI3K/Akt was shown to mediate a reduction of the activity of glycogen synthase kinase 3β (GSK-3β), which is involved in increased β-amyloid production and hyperphosphorylation of the tau protein [40].

Interestingly, the previously revealed antidiabetic activity of GSB-214 proved dependent on the activation of the PI3K/Akt pathway, as shown by a pharmacological inhibitory analysis [17]. Since it is well-known that AD and diabetes mellitus have a similar pathogenesis [18], this fact supports the idea that the PI3K/Akt pathway also contributes to the effects of GSB-214 in a streptozotocin-induced model reproducing all the major pathophysiological mechanisms of AD.

Figure 4 shows the putative mechanisms of action of GSB-214 in AD models. Additional studies are needed to identify the exact mechanisms of action of GSB-214 in an experimentally induced model of AD.

Activation of the PI3K/AKT signaling pathway by the dipeptide GSB-214, which had previously been identified in *in vitro* experiments [10], may promote neuroprotection by inhibiting pro-apoptotic proteins and activating anti-apoptotic proteins, as well as improve memory consolidation and, therefore, long-term memory through the activation of the regulator of mTOR protein synthesis.

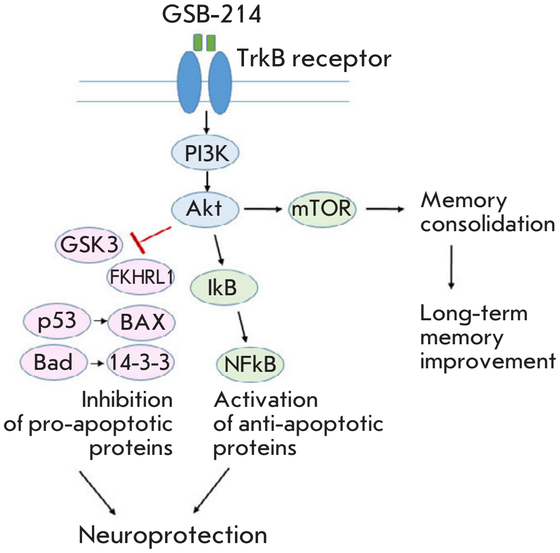

GSB-214
TrkB receptor
PI3K
Akt
mTOR
Memory consolidation
Long-term memory improvement
GSK3
FKHRL1
IkB
p53
BAX
Bad
14-3-3
NFkB
Inhibition of pro-apoptotic proteins
Activation of anti-apoptotic proteins
Neuroprotection

Fig. 4. The putative mechanisms of action of the BDNF dipeptide mimetic GSB-106 in AD models

CONCLUSIONS

Therefore, the low-molecular-weight BDNF mimetic GSB-214 dipeptide eliminates induced memory impairment in rats in the scopolamine- and streptozotocin-induced models of Alzheimer's disease. The effect of GSB-214 depends on the activation of Trk receptors. ●

REFERENCES

1. 2019 Alzheimer's disease facts and figures // Alzheimer's Dementia. 2019. V. 15. № 3. P. 321–387.
2. Giuffrida M.L., Copani A., Rizzarelli E. // Aging (Albany. NY). 2018. V. 10. № 8. P. 1791–1792.
3. Iulita M.F., Bistué Millón M.B., Pentz R., Aguilar L.F., Do Carmo S., Allard S., Michalski B., Wilson E.N., Ducatenzeiler A., Bruno M.A., et al. // Neurobiol. Dis. 2017. V. 108. P. 307–323.
4. Amidfar M., de Oliveira J., Kucharska E., Budni J., Kim Y.K. // Life Sci. 2020. V. 257. P. 118020.
5. Wang Z.H., Xiang J., Liu X., Yu S.P., Manfredsson F.P., Sandoval I.M., Wu S., Wang J.Z., Ye K. // Cell Rep. 2019. V. 28. № 3. P. 655.
6. Arancibia S., Silhol M., Moulière F., Meffre J., Höllinger I., Maurice T., Tapia-Arancibia L. // Neurobiol. Dis. 2008. V. 31. № 3. P. 316–326.
7. Nagahara A.H., Mateling M., Kovacs I., Wang L., Eggert S., Rockenstein E., Koo E.H., Masliah E., Tuszynski M.H. // J. Neurosci. 2013. V. 33. № 39. P. 15596–15602.
8. Jiao S.S., Shen L.L., Zhu C., Bu X.L., Liu Y.H., Liu C.H., Yao X.Q., Zhang L.L., Zhou H.D., Walker D.G., et al. // Transl. Psychiatry. 2016. V. 6. № 10. P. e907.
9. Kopec B., Zhao L., Rosa-Molinar E., Siahaan T. // Med. Res. Arch. 2020. V. 8. № 2. P. 2043.
10. Gudasheva T.A., Povarnina P.Y., Tarasiuk A.V., Seredenin S.B. // Med. Res. Rev. 2021. № 41. P. 2746–2774.
11. Longo F.M., Massa S.M. // Nat. Rev. Drug Discov. 2013. V.12. №7. P.507–525.
12. Zhang Z., Liu X., Schroeder J.P., Chan C.-B., Song M., Yu S.P., Weinshenker D., Ye K. // Neuropsychopharmacology. 2014. V. 39. № 3. P. 638–650.
13. Aytan N., Choi J.K., Carreras I., Crabtree L., Nguyen B., Lehar M., Blusztajn J.K., Jenkins B.G., Dedeoglu A. // Eur. J. Pharmacol. 2018. V. 828. P. 9.
14. Bollen E., Vanmierlo T., Akkerman S., Wouters C., Steinbusch H.M.W., Prickaerts J. // Behav. Brain Res. 2013. V. 257. P. 8–12.
15. Gudasheva T.A., Tarasyuk A.V., Pomogaibo S.V., Logvinov I.O., Povarnina P.Yu., Antipova T.A., Seredenin S.B. // Russ. J. Bioorganic Chem. 2012. V. 38. № 3. P. 280–290.
16. Gudasheva T.A., Povarnina P., Logvinov I.O., Antipova T.A., Seredenin S.B. // Drug Des. Devel. Ther. 2016. V. 10. P. 3545–3553.
17. Yagubova S.S., Ostrovskaya R.U., Gudasheva T.A., Seredenin S.B. // Bull. Exp. Biol. Med. 2020. V. 169. № 6. P. 712–715.
18. de la Monte S.M., Wands J.R. // J. Diabetes Sci. Technol. 2008. V. 2. № 6. P. 1101.
19. Ennaceur A., Delacour J. // Behav. Brain Res. 1988.

V. 31. № 1. P. 47–59.

20. Antunes M., Biala G. // Cogn. Process. 2012. V. 13. № 2. P. 93–110.

21. Beldjoud H., Barsegyan A., Roozendaal B. // Front. Behav. Neurosci. 2015. V. 9. P. 108.

22. Volkova A.A., Povarnina P.Yu., Nikiforov D.M., Gudasheva T.A., Seredenin S.B.// Pharm. Chem. J. 2022. V. 56. № 4. P. 3–6.

23. Richter N., Beckers N., Onur O.A., Dietlein M., Tittgemeyer M., Kracht L., Neumaier B., Fink G.R., Kukolja J. // Brain. 2018. V. 141. № 3. P. 903–915.

24. Ostrovskaya R.U., Mirzoev T.Kh., Firova F.A. // Experimental and Clinical Pharmacology. 2001. V. 64. № 2. P. 11–14.

25. van Dam D., De Deyn P.P. // Nat. Rev. Drug Discov. 2006. V. 5. № 11. P. 956–970.

26. Bhuvanendran S., Kumari Y., Othman I., Shaikh M.F. // Front. Pharmacol. 2018. V. 9. P. 665.

27. Rai S., Kamat P.K., Nath C., Shukla R. // J. Neuroimmunol. 2013. V. 254. № 1–2. P. 1–9.

28. Kamat P.K., Kalani A., Rai S., Tota S.K., Kumar A., Ahmad A.S. // Mol. Neurobiol. 2016. V. 53. № 7. P. 4548–4562. https://link.springer.com/article/10.1007/s12035-015-9384-y.

29. Kamat P.K. // Neural Regen. Res. 2015. V. 10. № 7. P. 1050.

30. Salkovic-Petrisic M., Hoyer S. // J. Neural Transm. Suppl. 2007. № 72. P. 217–233.

31. Ravelli K.G., Rosário B. dos A., Camarini R., Hernandes M.S., Britto L.R. // Neurotox. Res. 2017. V. 31. № 3. P. 327–333.

32. Bassani T.B., Turnes J.M., Moura E.L.R., Bonato J.M., Cóppola-Segovia V., Zanata S.M., Oliveira R.M.M.W., Vital M.A.B.F. // Behav. Brain Res. 2017. V. 335. P. 41–54.

33. Afshar S., Shahidi S., Rohani A.H., Komaki A., Asl S.S. // Psychopharmacol. 2018. V. 235. № 10. P. 2809–2822.

34. Mugwagwa A.T., Gadaga L.L., Pote W., Tagwireyi D. // J. Neurodegener. Dis. 2015. V. 2015. P. 1–9.

35. Switon K., Kotulska K., Janusz-Kaminska A., Zmorzynska J., Jaworski J. // Neuroscience. 2017. V. 341. P. 112–153.

36. Hernandez P.J., Abel T. // Neurobiol. Learn Mem. 2008. V. 89. № 3. P. 293–311.

37. Jobim P.F.C., Pedroso T.R., Werenicz A., Christoff R.R., Maurmann N., Reolon G.K., Schröder N., Roesler R. // Behav. Brain Res. 2012. V. 228. № 1. P. 151–158.

38. Porsteinsson A.P., Isaacson R.S., Knox S., Sabbagh M.N., Rubino I. // J. Prev. Alzheimer's Dis. 2021. V. 8. № 3. P. 371–386.

39. Reichardt L.F. // Philos. Trans. R. Soc. B Biol. Sci. 2006. V. 361. № 1473. P. 1545–1564.

40. Long H.Z., Cheng Y., Zhou Z.W., Luo H.Y., Wen D.D., Gao L.C. // Front. Pharmacol. 2021. V. 12. P. 648636.

The Fallout of Catastrophic Technogenic Emissions of Toxic Gases Can Negatively Affect Covid-19 Clinical Course

G. Succi[1], W. Pedrycz[2], A. P. Bogachuk[3], A. G. Tormasov[1], A. A. Belogurov[3,4], A. Spallone[3,5*]
[1]Innopolis University, Innopolis, 420500 Russia
[2]University of Alberta, Edmonton (AB), T6G 2R3 Canada
[3]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, 117997 Russia
[4]Evdokimov Moscow State University of Medicine and Dentistry, Department of Biological Chemistry, Moscow, 127473 Russia
[5]Neurological Centre of Latium, Institute of Neurological Sciences, Rome, 00178 Italy
[*]E-mail: aldospallone@hotmail.com

Received June 16, 2022; in final form, November 29, 2022
DOI: 10.32607/actanaturae.11754

ABSTRACT **The coronavirus D-19 (Covid-19) pandemic has shaken almost every country in the world: as we stand, 6,3 million deaths from the infection have already been recorded, 167,000 and 380,000 of which are in Italy and the Russian Federation, respectively. In the first wave of the pandemic, Italy suffered an abnormally high death toll. A detailed analysis of available epidemiological data suggests that that rate was shockingly high in the Northern regions and in Lombardy, in particular, whilst in the southern region the situation was less dire. This inexplicably high mortality rate in conditions of a very well-developed health care system such as the one in Lombardy – recognized as one of the best in Italy – certainly cries for a convincing explanation. In 1976, the small city of Seveso, Lombardy, experienced a release of dioxin into the atmosphere after a massive technogenic accident. The immediate effects of the industrial disaster did not become apparent until a surge in the number of tumors in the affected population in the subsequent years. In this paper, we endeavor to prove our hypothesis that the release of dioxin was a negative cofactor that contributed to a worsening of the clinical course of COVID-19 in Lombardy.**

KEYWORDS **SARS-CoV-2, COVID-19, Italy, Seveso, dioxin.**

ABBREVIATIONS **COVID-19 – COronaVIrus Disease 2019; PM-10 – Particulate Matter-10 Microns or less; ISTAT – Italian National Institute of Statistics; ISPRA – Istituto Superiore per la Protezione e la Ricerca Ambientale/Italian Superior Institute for Protection and Environmental Research.**

INTRODUCTION

The coronavirus D-19 (Covid-19) pandemic has hit almost every country in the world as it has spread west from China to Europe, the U.S., and later South America, Africa, and the Russian Federation. At this juncture, 6.3 million COVID-19 deaths have been reported worldwide, with 380,000 of those in the Russian Federation alone. The impact of the pandemic has been particularly severe in several European countries, such as Spain, France, Belgium, the UK, and Italy, whilst other countries such as Portugal, Germany, the Scandinavian states, and Eastern Europe, in general, have had it relatively easy. Italy, in particular, has had a shockingly high mortality rate, one that significantly exceeded the death rate observed in the rest of the world.

But a closer look at the epidemiological data would suggest that this high rate was mainly concentrated just in the Northern regions, in Lombardy in particular, whilst in the southern region the clinical course of most patients was more favorable, as our group had predicted well in advance [1]. This situation is particularly troubling if the general mortality rate is compared with the one that prevailed in the previous five years. This inexplicably high mortality rate in the context of a very well-developed health care system such as the one in Lombardy – largely recognized as one of the best, if not the best, in the coun-

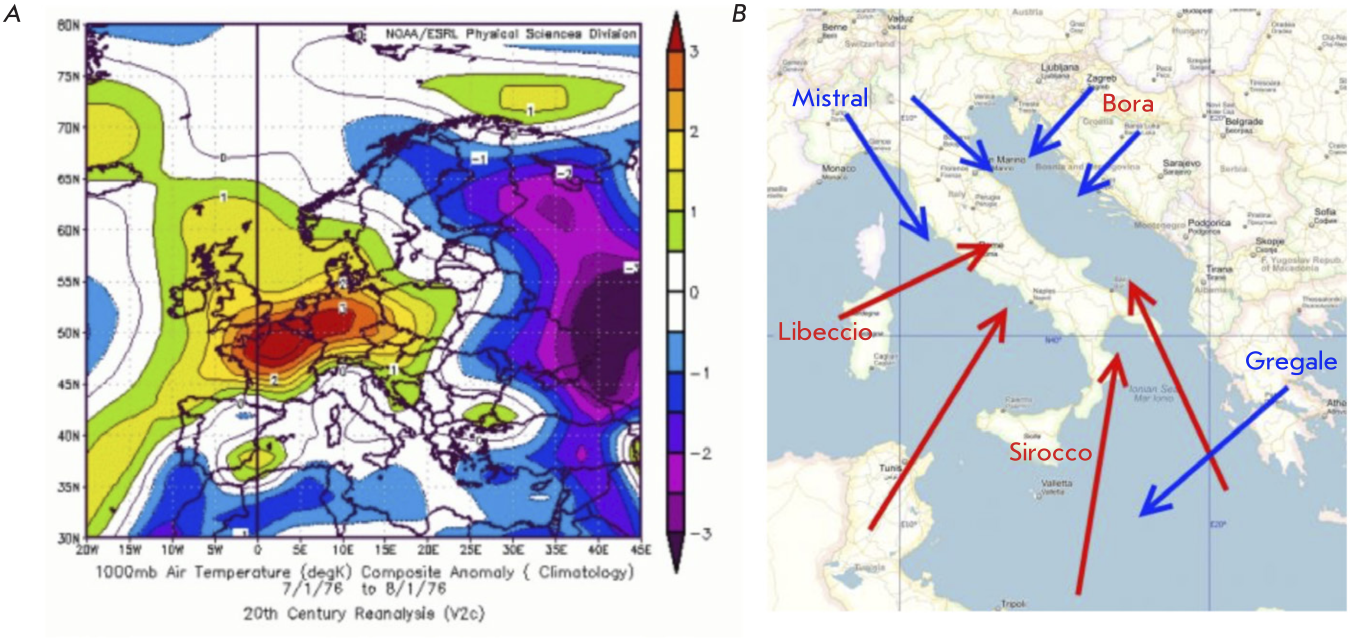

A
NOAA/ESRL Physical Sciences Division
80N
75N
70N
65N
60N
55N
50N
45N
40N
35N
30N
20W 15W 10W 5W 0 5E 10E 15E 20E 25E 30E 35E 40E 45E
3
2
1
0
−1
−2
−3
1000mb Air Temperature (degK) Composite Anomaly (Climatology)
7/1/76 to 8/1/76
20th Century Reanalysis (V2c)
B
Mistral
Bora
Libeccio
Gregale
Sirocco

Fig. 1. (*A*) Typical weather conditions in Northern Italy, around the Alps https://progettoscienze.com/2016/09/29/i-grandi-classici-della-scienza-libellus-de-ratiociniis-in-ludo-alee/#jp-carousel-6901 (*B*) Typical structure of winds in Italy http://sailroad.ru/article/lociya-srednej-dalmacii-chast-2

try [2] – certainly calls for a satisfactory explanation. Some experts have directed their attention at the potential negative role of PM-10, which are overrepresented in Lombardy [3] and some neighboring regions also significantly affected by the Covid-19 pandemic. However, if we assume that this hypothesis is sound, it becomes hard to explain why California, which is highly polluted and seriously affected by PM-10, appeared definitely less affected than other states, with New York first in mind, where the air concentration of PM-10 is lower.

In 1976 the small city of Seveso, which is relatively close to Brescia, Bergamo and Milan, became sadly known in the world for an accidental escape of dioxin. The immediate effects were mild, but that was before an increased number of tumors began to appear in the affected population in subsequent years [4]. In our work, we hypothesize that the gas escape had a negative cofactor role in the worse clinical course of the Covid-19 pandemic for patients in Lombardy.

EXPERIMENTAL

We conducted a study correlating the distance from the epicenter of the escape of dioxin, Seveso, to the rate of mortality of the potentially affected provinces. We studied the local mortality rate from the Covid-19 infection as a percentage of the dead vs. infected patients and compared data for Lombardy with those for other world regions where a dramatic leak of toxic gases had occurred: the city of Bhopal (India), where a significant accidental toxic gas release occurred in 1980 from a local Union Carbide factory, something that was considered at the time as the worst industrial disaster in history [5, 6].

We retrieved data about the local weather conditions and winds directions at the time of the accidents and also calculated the distance between the sites of the escapes and those most clinically affected. We also analyzed the air pollution of the three sites as measured by the PM-10 concentration. The gases were in fact different: 2,3,7,8-tetrachlorodibenzodioxin in the case of Seveso [7] and methyl isocyanate in the case of Bhopal [8–10]. Nevertheless, both gases are known to be mutagenic and cancerogenic [7, 11–15].

In both scenarios the analysis of local weather conditions allowed us to somewhat reconstruct the possible spread of the escaped gases on account of the effects of the winds. In the Seveso case, nice weather conditions in Lombardy, together with high pressure in the Alps (*Fig. 1A*), favored the Mistral pushing the gases south east; i.e., in the direction of Bergamo, Brescia and further south up to the western provinces of Veneto and Emilia Romagna. Some other components of the Mistral could also have pushed the gas towards eastern Piemonte and the northern part of Liguria (*Fig. 1B*). In the Bhopal region, where the ac-

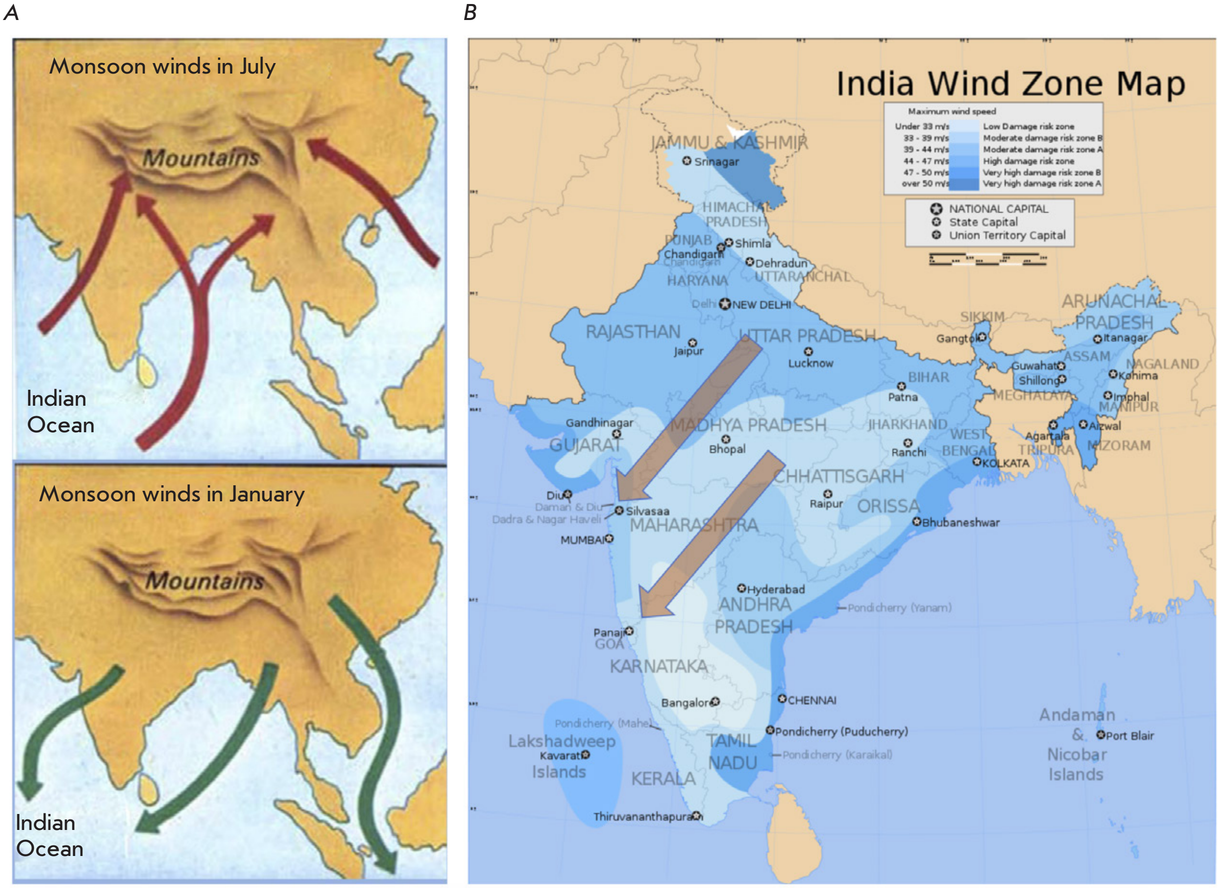


Fig. 2. (*A*) Typical structure of Winds in India https://cloud.prezentacii.org/19/04/142027/images/screen7.jpg (*B*) Winds around Bhopal https://commons.wikimedia.org/wiki/File:India_wind_zone_map_en.svg

cident occurred in December, the Monsoons typically flow from north to the southwest (*Fig. 2A*): so, the gases escape would have spread from the Bhopal region of Madhya Pradesh to the neighboring state of Maharashtra (*Fig. 2B*). As far as Italy was concerned, we also considered the possible impact of Chinese immigrants, as well as the course of the infection vs. the density of the local population.

DATA SOURCES

We used data available in several public databases. *Table 1* shows the distribution of the 2020 death rate compared with previous years as reported by ISTAT, the Italian Institute of statistics. The increase is particularly notable in the north, specifically in Lombardy [1]. Graphic is reported in *Fig. 3*. Data from Chinese immigration in Italy are also from ISTAT, which also provided data on the local population. Data on the infection incidence (https://github.com/pcm-dpc/COVID-19) come from the official GitHub repository of the Italian government [16] and is represented in *Fig. 4*. The data on infections and the death rate in India come from publicly available sources and reports of the death rate in the potentially affected regions of India as compared to the rest of the country (*Table 2*). Data regarding the PM-10 concentration in Italy were retrieved from the repository of ISPRA (Istituto Superiore per la Protezione e la Ricerca Ambientale/Italian Superior Institute for Protection and Environmental Research) [17]. *Figure 5* shows the level of PM-10 concentration in Europe. It shows how its concentration was increased over the Padana landscape [18].

STATISTICAL ANALYSIS

To analyze the data, we used the non-parametric Spearman's Rank correlation coefficient [19]. Such a method includes no assumption on the underlying data, apart from being at least on an ordinal scale, which is always the case in our analyses. As the threshold of significance, we considered an α-level of 0.05, as customary. In the case where multiple

Table 1. Variation of the % of deaths in the period under consideration (Feb. 15 – Apr. 15) with respect to the same period in 2019

Province	Variation, %
Agrigento	-22.22
Cagliari	-16.67
Matera	-7.69
Crotone	-6.45
Catania	-5.56
Roma	3.94
Perugia	5.15
Arezzo	6.60
Lecce	6.72
Vibo Valentia	7.14
Ravenna	7.30
Foggia	9.16
Taranto	9.46
Messina	10.81
Sassari	10.93
Catanzaro	11.11
Teramo	11.11
Potenza	11.49
Ferrara	12.28
Salerno	12.98
Barletta-Andria-Trani	13.27
Palermo	14.04
Pisa	14.12
Siena	14.17
Fermo	15.00
Belluno	15.65
Venezia	17.41
Napoli	17.57
Brindisi	17.90
Trapani	18.95
Bologna	19.02
Macerata	19.32
Verona	19.47
Terni	20.00
Bari	20.15
Livorno	20.29
Forli-Cesena	21.33
Grosseto	21.54
Lucca	21.62
Rovigo	22.77
Oristano	23.58
Varese	24.33
Frosinone	24.49
Genova	24.65
Pistoia	24.79
Caltanissetta	25.00
Ascoli Piceno	25.53
Savona	25.61
Asti	26.46
La Spezia	26.55
Como	26.59
Torino	26.88
Pescara	26.88
Modena	27.51
Firenze	27.66
L'Aquila	28.00
Padova	28.03
Cosenza	28.05
Reggio di Calabria	28.26
Viterbo	28.38
Ancona	28.68
Massa Carrara	28.84
Vicenza	28.92
Verbano-Cusio-Ossola	28.99
Udine	29.41
Cuneo	30.02
Imperia	31.17
Nuoro	31.73
Treviso	31.88
Sondrio	32.34
Pordenone	33.33
Milano	33.80
Novara	34.73
Rimini	34.85
Chieti	36.36
Gorizia	36.84
Vercelli	37.61
Avellino	38.20
Monza e Brianza	38.86
Siracusa	39.34
Sud Sardegna	39.58
Alessandria	39.64
Latina	40.00
Isernia	40.00
Campobasso	40.82
Benevento	41.67
Trento	42.72
Reggio nell'Emilia	43.48
Mantova	43.77
Enna	44.78
Biella	45.48
Aosta	47.65
Pesaro e Urbino	49.56
Lecco	50.17
Pavia	50.51
Ragusa	51.85
Parma	56.97
Caserta	59.26
Brescia	64.25
Piacenza	68.57
Lodi	70.13
Cremona	71.93
Bergamo	78.77

Note: the province of Bolzano is not reported.

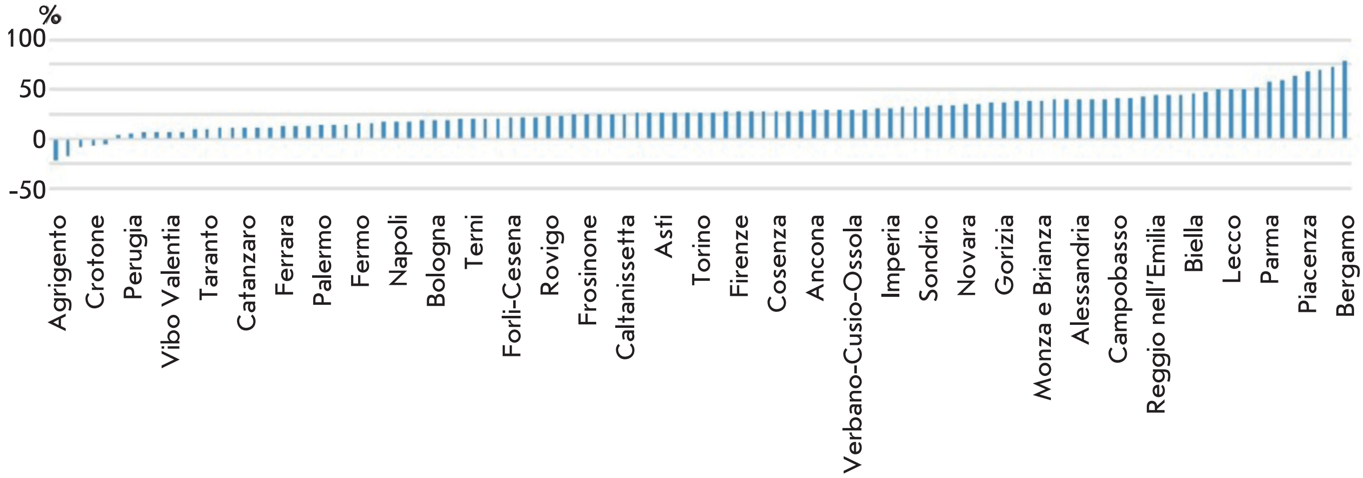
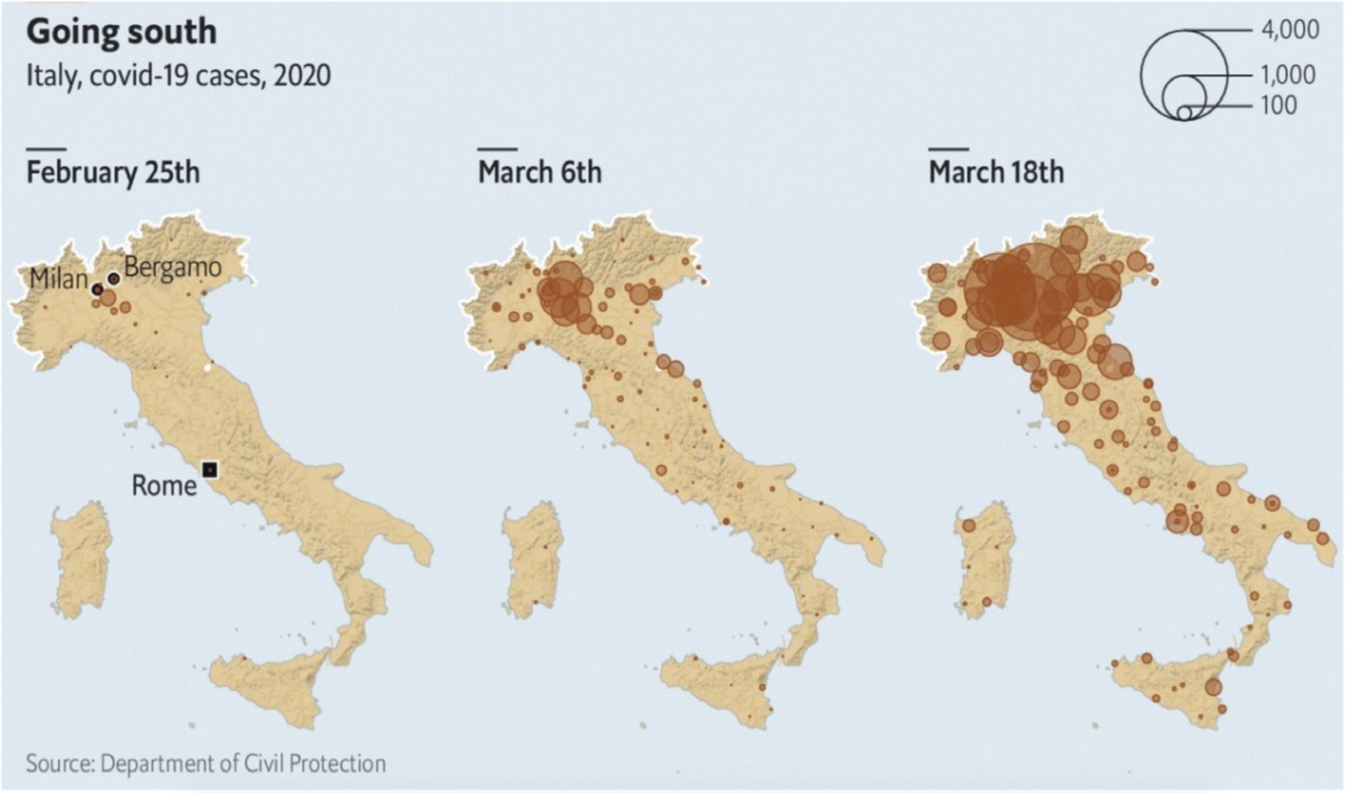


Fig. 3. Percentage of variation of deaths across the provinces of Italy in the period under consideration (February 15 – April 15) – We present the extremes; the details are in *Table 1*



Source: Department of Civil Protection

Fig. 4. Covid-19 map spread in Italy. https://www.economist.com/europe/2020/03/19/italy-is-overtaking-china-as-the-country-worst-hit-by-covid-19

hypotheses were being considered, we applied the Bonferroni correction [20]. In the case of an analysis of multiple factors, we used ANOVA [21]: again, considering the α-level mentioned above.

RESULTS

Chinese Immigration

The presence of immigrants from China in Italy (https://www.tuttitalia.it/statistiche/cittadini-stranieri/repubblica-popolare-cinese/) is not a factor in the spread of the virus [22]. In fact, in 2019, the number of Chinese present in Milan was 40,438 (1.25% of the total population), whilst in Rome their number was 22,815 (0,52%). The provinces with the highest percentage of increase in deaths had the following numbers: Bergamo 4,488 (0.40%), Cremona 1,362 (0.35%), and Lodi 757 (0.33%).

Population Density

Social proximity does not appear to affect the contagion and the death rate in Italy. We found no significant nonparametric correlation between the density of the population and the increase in mortality with respect to the last five years average (the p values are 0.083 and 0.071 respectively, indeed not significant), or between density and infection spread (0.17; again, absolutely not significant).

Influence of the PM-10 level

The PM-10 appears to have an effect considering the number of days above the threshold in Italy and, in particular, in Lombardy. There is a correlation of 0.40 with the number of deaths in 2019 ($p < 10^{-4}$) and a correlation of 0.38 with the 5-year average of number of deaths ($p < 10^{-3}$). There is also a correlation of 0.41 with the percentage of infected people ($p < 10^{-4}$). However, if we consider together the effects of the distance from Seveso and the presence of PM-10 in a ranked ANOVA, we observe that the distance from Seveso retains its significance ($t = -15.57$, $p < 10^{-8}$), while the presence of PM-10 does not.

Table 2: Data about Covid-19 mortality in India. https://www.mohfw.gov.in/

S. No.	Name of State/ UT	Total Confirmed cases	Cured/ Discharged/ Migrated	Deaths
1	Andaman and Nicobar Islands	33	33	0
2	Andhra Pradesh	2407	1456	50
3	Arunachal Pradesh	1	1	0
4	Assam	101	41	2
5	Bihar	1262	475	8
6	Chandigarh	191	51	3
7	Chhattisgarh	86	59	0
8	Dadar Nagar Haveli	1	0	0
9	Delhi	10054	4485	160
10	Goa	29	7	0
11	Gujarat	11379	4499	659
12	Haryana	910	562	14
13	Himachal Pradesh	80	44	3
14	Jammu and Kashmir	1183	575	13
15	Jharkhand	223	113	3
16	Karnataka	1147	509	37
17	Kerala	601	497	4
18	Ladakh	43	24	0
19	Madhya Pradesh	4977	2403	248
20	Maharashtra	33053	7688	1198
21	Manipur	7	2	0
22	Meghalaya	13	11	1
23	Mizoram	1	1	0
24	Odisha	828	220	4
25	Puducherry	13	9	1
26	Punjab	1964	1366	35
27	Rajasthan	5202	2992	131
28	Tamil Nadu	11224	4172	78
29	Telengana	1551	992	34
30	Tripura	167	85	0
31	Uttarakhand	92	52	1
32	Uttar Pradesh	4259	2441	104
33	West Bengal	2677	959	238
Total number of confirmed cases in India		96169	36824	3029

Distance from Seveso and Bhopal

The distance from Seveso appears to be a determining factor (*Fig. 3*). In terms of increase in deaths with respect to 2019 we found a very strong correlation: -0.82 ($p < 10^{-24}$), whilst with respect to the average of the last five years it was -0.83, ($p < 10^{-25}$). In terms of the percentage of infected population, the correlation is even higher, at -0.88 ($p < 10^{-32}$). In summary, the closer to Seveso the analyzed sites were, the higher the rate of infected population and Covid-19-related deaths were. In India, the correlation be-

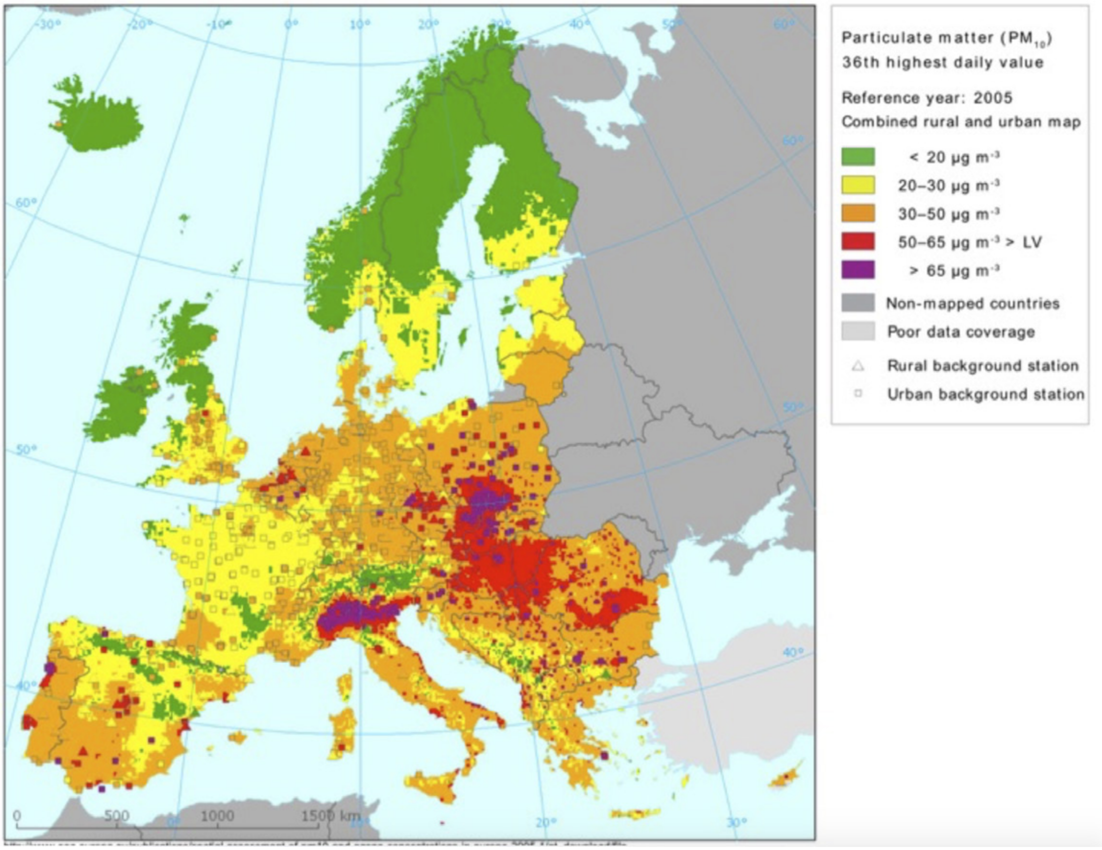


Fig. 5. PM-10 contamination in Europe https://commons.wikimedia.org/wiki/File:PM10_in_Europe.png

tween distances from Bhopal is significant on both the reported percentage of deaths due to coronavirus (-0.52, $p < 0.01$) and of infected people (-0.36, $p < 0.05$) (*Table 3*).

DISCUSSION

The possibility of a toxic gas escape that occurred 40 years ago playing a role in the increased incidence of complicated clinical courses in the recent Covid-19 infection is an intriguing, albeit difficult to demonstrate, hypothesis. As a result of both accidents, two different toxic gases were released, but both gases were characterized by high carcinogenicity [7, 11–15, 23, 24]. An increased mortality rate from COVID-19 was observed in all regions potentially exposed to the gases spread by the winds prevailing at the time of the accident.

An increased mortality rate due to the Covid-19 infection was witnessed in all the regions potentially touched by the gas leaks. This is also intimated in the observation of the possible effects of the winds active in those particular times of the year. This death rate increase was particularly striking in Lombardy, a fact that continues to require a plausible explanation. The particularly high virulence of the virus that affected the North of Italy was claimed as a possible reason for the high death toll [1]. Even if we assume that the better clinical course observed in the southern Italian regions was the result of heeding the lessons learned when the disease coursed through the northern parts of the country, the mortality rate difference remains hard to explain.

The possible detrimental effects of the PM-10 pollution has been invoked as a negative factor that has aided a more aggressive clinical course of the epidemic due to its chronic irritative impact on the respiratory system [25]. However, as we noted above, this hypothesis is somewhat contradicted by the observation that the impact of the epidemic in California has been definitely milder than it has been in New York, although the air concentration of PM-10 is much higher in California [26]. So, the detrimental effect of PM-10 pollution cannot be the sole reason for what was observed in Lombardy.

Other claims refer to the presence of immigrants from China. The available data from ISTAT show that on January 1, 2019, the number of Chinese present in Milan was higher than that in Rome. However, the

Table 3. Revealed statistically significant patterns of the significance of factors affecting the level of infection and mortality from COVID-19

Country	Possible cause	Possible effect, %	Spearman's Rank Correlation
Italy	Distance from Seveso	Variation of death over 2019 Variation of death over 5-year average Infection	-0.82 ($p < 10^{-24}$) -0.83 ($p < 10^{-25}$) -0.88 ($p < 10^{-32}$)
	Number of days of PM-10 over threshold	Variation of death over 2019 Variation of death over 5-year average Infection	0.40 ($p < 10^{-4}$) 0.38 ($p < 10^{-3}$) 0.41 ($p < 10^{-4}$)
India	Distance from Bhopal	Deaths due to COVID-19 Infection	-0.52 ($p < 10^{-2}$) -0.36 ($p < 0.05$)

provinces with the highest increase in deaths had lower numbers of Chinese immigrants. There are also claims that social proximity increases the contagion rate and, consequently, the death rate. However, we found that density did not push the mortality rate upward as relates to 2019 and to the last five years. Also, if we consider the number of infections, in this case there is also no significant correlation.

We hypothesize that the fallout of the Seveso accident – perhaps in addition to the detrimental effects of air pollution – would have acted synergically in Lombardy to make the clinical course of the coronaviral infection there particularly aggressive. It may have acted not only by predisposing residents, as a consequence of air pollution's effect on the respiratory system of Lombard patients, to viral attacks, in particular to a significantly more aggressive course of the autoimmune reaction towards the alveoli the virus induces, but also through some gene-modifying mechanism that had taken place during the preceding 45 years and acted somehow by reinforcing the aforementioned autoimmune process. The other case, Bhopal, India, experienced an increased mortality rate as compared to the rest of the country. However, this difference, albeit significant, was not as striking as the one observed in Lombardy. We would venture that, in the region of Bhopal, the air concentration of PM-10 is not as significant as it is in the Padana landscape, which is a well-known site of significantly polluted air.

To support our claims, we used the robust Spearman's Rank correlation coefficient. We considered first the relationship between the unequivocal number of variation of deaths in relationship to the previous years. The resulting value of the correlation of the distance from Seveso and the increase in deaths with respect to 2019 is impressive (-0.82, $p < 10^{-24}$), and it is even more impressive with respect to the average for the last five years (-0.83, $p < 10^{-25}$). We have also considered the relationship between distance to Seveso and the percentage of the infected population, and in this case the correlation is even higher -0.88 ($p < 10^{-32}$).

For conclusiveness, we have also considered the claimed effect of PM-10, particularly by calculating the number of days above the safe threshold. We have noticed that, indeed, there is an impact, by far below the one related to the distance from Seveso (0.40, $p < 10^{-4}$). The average number of deaths over 5 years (0.38, $p < 10^{-3}$) and the percentage of those infected with COVID-19 (0.41, $p < 10^{-4}$) also correlated with elevated levels of PM-10. We have built a ranked ANOVA to attempt to determine the joint contribution of the number of days of PM-10 above the threshold and the distance from Seveso. In performing such an analysis, we arrived at the conclusion that the distance from Seveso remains highly statistically significant, while the number of days of PM-10 above the threshold completely loses such significance.

To bolster our hypothesis, we turned our attention to the case of India, where we directly tested the presence of a correlation between the distance from Bhopal and the reported rate of infected and dead people from Coronavirus. The historical data on the total number of deaths and on the presence of PM-10 was not available to us. So, we had to rely only on the public data specific to the disease in 2020. In this case, we also found a statistically significant Spearman's rank correlation between the distance from Bhopal and the percentage of infected people (-0.36, $p < 0.05$), as well as that of dead people (-0.52, $p < 0.01$).

CONCLUSIONS

Our hypothesis, obviously, requires confirmation, perhaps through a study comparing certain genomic characteristics of Lombardy longtime residents with those of relatively recent immigrants. As a matter of fact, a strikingly low presence of immigrants amongst the Covid-19 patients admitted to the ICUs of Lombardy hospitals has been observed [27], and a convincing explanation of that fact has yet to be provided. At the same time, we could not find in the scientific literature and statistical data direct evidence of increased mortality with seasonal influenza diseases in that region until the spring of 2020.

The technogenic catastrophe and the complicated course of COVID-19 in Lombardy may have something to do with the increased level of diabetes mellitus, oncological, and autoimmune disorders. Thus, population studies of mortality for 25 years since the accident in 1976, conducted by Consonni and colleagues, revealed increased additional mortality from diabetes among women in all areas of pollution, dependent on the degree of damage to the area [28]. According to available data, during the first 25 years after the technogenic accident (1976–2001), no increase in the total cancer mortality was detected throughout the affected areas. However, once the mortality rate was studied some 20 plus years after the explosion, an increase in cancer mortality was recorded in the area with the most severe pollution [28]. A similar correlation was observed with autoimmune diseases. In the affected areas, an inverse correlation was found between the level of immunoglobulin and dioxin in the blood plasma of adult patients [29]. At the same time, another study found an increase in the titers of antinuclear antibodies, an increase in the deposition of immune complexes, and a decrease in the number of natural killers in patients from the affected areas [30].

The half-life of dioxin in the body is 7–11 years. Since the disaster in Seveso occurred in 1976, the direct effect of dioxin can no longer be taken into account. Nevertheless, it is interesting to study the delayed effects of this substance on the human body. Since this manuscript considers the possible connection between residents of this particular area and the higher mortality rate from COVID-19, a next stage of this study could be the inclusion in the study sample of only the generation of people who directly experienced the accident of 1976 or moved to Seveso for 7–11 years until the half-life of dioxin expired. In a separate comparison group, it is possible to include the descendants of people who were affected by the accident and stayed in the territory. It is especially interesting to follow the individuals who survived the accident and their descendants who left for other regions of Italy and also suffered the new coronavirus infection. Unfortunately, at the moment (since the study is retrospective), such information is not available. Moreover, such data are not present either in open statistical data or in outpatient records. Therefore, a much larger resource is required for its systematization.

By focusing future research on the genomics and proteomics of affected patients in the area of technogenic disasters, especially young patients with a severe clinical course, it is possible not only to test the validity of our hypothesis, but also to predict the genetic determinants of individuals with a potentially worse prognosis of COVID-19. Such data could make the approach to treatment of COVID-19 more personalized, as well as identify risk groups that must be prioritized regarding vaccination, revaccination, and protection in terms of limiting social contacts. ●

This work was carried out within the framework of the Russian Ministry of Education and Science project No. 075-15-2021-1049.

REFERENCES

1.Masyagin S., Mazzara M., Succi G., Spallone A., Volpi A. // Biomed. J. Sci. Tech. Res. 2020. V. 27. № 5. P. 21056–21062.

2. Lamberti-Castronuovo A., Parotto E., Della Corte F., Hubloue I., Ragazzoni L., Valente M. // Front. Public Hlth. 2022. V. 10. P. 1034196. doi: 10.3389/fpubh.2022.1034196.

3. Carugnoa M., Consonni D., Bertazzi P.A., Biggeric A., Baccini M. // Environ. Pollution. 2017. V. 227. P. 280–286.

4. Pesatori A.C., Consonni D., Rubagotti M., Grillo P., Bertazzi P.A. // Environ. Hlth. 2009. V. 8. P. 39. https://doi.org/10.1186/1476-069X-8-39.

5. De S., Shanmugasundaram D., Singh S., Banerjee N., Soni K.K., Galgalekar R. // Public Hlth. 2020. V. 186. P. 20–27. doi: 10.1016/j.puhe.2020.06.043.

6. Varma R., Varma D.R. // Bull. Sci. Technol. Soc. 2005. V. 25. № 1. P. 37–45.

7. Homberger E., Reggiani G., Sambeth J., Wipf H.K. // Ann. Occup. Hyg. 1979. V. 22. № 4. P. 327–370.

8. Senthilkumar C.S., Malla T.M., Akhter S., Sah N.K., Ganesh N. // Cien. Saude Colet. 2020. V. 25 (suppl 2). P. 4225–4230. doi: 10.1590/1413-812320202510.2.28682020.

9. The public health implications of the Bhopal disaster. Report to the Program Development Board, American Public Health Association. Bhopal Working Group // Am. J. Public Hlth. 1987. V. 77. № 2. P. 230–236. doi: 10.2105/ajph.77.2.230.

10. Dourson M.L., Kohrman-Vincent M.J., Allen B.C. // Regul. Toxicol. Pharmacol. 2010. V. 58. № 2. P. 181–188. doi: 10.1016/j.yrtph.2010.04.006.

11. Ingrid E. // Univ. Press (India). 2005. doi: 10.13140/2.1.3457.5364.

12. Kassie F., Laky B., Nobis E., Kundi M., Knasmüller S. // Mutat. Res. 2001. V. 492. № 1–2. P. 111–113. doi: 10.1016/s1383- 5718(00)00140-6.

13. Kobets T., Smith B.P.C., Williams G.M. // Foods. 2022.

V. 11. № 18. P. 2828. doi: 10.3390/foods11182828.
14. Eskenazi B., Mocarelli P., Warner M., Needham L., Patterson D.G. Jr, Samuels S., Turner W., Gerthoux P.M., Brambilla P. // Environ. Hlth Perspect. 2004. V. 112. № 1. P. 227.
15. Vuong T.P. // Toxics. 2022. V. 10. № 7. P. 384. doi: 10.3390/toxics10070384.
16. Monzani D., Vergani L., Pizzoli S.F.M., Marton G., Pravettoni G. // J. Med. Internet. Res. 2021. V. 23. № 10. e29820. doi: 10.2196/29820.
17. Mannocci A., Ciarlo I., D'Egidio V., Del Cimmuto A., de Giusti M., Villari P., La Torre G. // J. Environ. Public Hlth. 2019. V. 2019. P. 2058467. doi: 10.1155/2019/2058467.
18. Farahani V.J., Altuwayjiri A., Pirhadi M., Verma V., Ruprecht A.A., Diapouli E., Eleftheriadis K., Sioutas C. // Environ. Sci. Atmos. 2022. V. 2. № 5. P. 1076–1086. doi: 10.1039/d2ea00043a.
19. Davies S.M., Geppert J., McClellan M., McDonald K.M., Romano P.S., Shojania K.G. Refinement of the HCUP Quality Indicators. Rockville (MD): Agency for Healthcare Research and Quality (US); 2001 May. Report No.: 01-0035.
20. Dunn O.J. // J. Amer. Statist. Ass. 1961. V. 56. № 293. P. 52–64.
21. Rayner J.C.W., Best D.J. // Australian New Zealand J. Statistics. 2013. V. 55. № 3. P. 305–319.
22. McKee M., Gugushvili A., Koltai J., Stuckler D. // Int. J. Hlth Policy Manag. 2021. V. 10. № 8. P. 511–515. doi: 10.34172/ijhpm.2020.124.
23. Benoit L., Koual M., Tomkiewicz C., Bats A.S., Antignac J.P., Coumoul X., Barouki R., Cano-Sancho G. // Environ. Int. 2022. V. 170. P. 107615. doi: 10.1016/j.envint.2022.107615.
24. O'Malley M., Barry T., Verder-Carlos M., Rubin A. // Am. J. Ind. Med. 2004. V. 46. № 1. P. 1–15. doi: 10.1002/ajim.20037. PMID: 15202120.
25. Colucci M.E., Veronesi L., Roveda A.M., Marangio E., Sansebastiano G. // Ig Sanita Publ. 2006. V. 62. № 3. P. 289–304.
26. Sannigrahi S., Pilla F., Maiti A., Bar S., Bhatt S., Kaparwan A., Zhang Q., Keesstra S., Cerda A. // Environ. Res. 2022. V. 210. P. 112818. doi: 10.1016/j.envres.2022.112818.
27. Grasselli G., Zangrillo A., Zanella A., Antonelli M., Cabrini L., Castelli A., Cereda D., Colucello A., Foti G., Fumagalli R., et al. // Am. J. Med. Assoc. 2020. V. 323. № 16. P. 1574–1581.
28. Consonni D., Pesatori A.C., Zocchetti C., Sindaco R., D'Oro L.C., Rubagotti M., Bertazzi P.A. // Am. J. Epidemiol. 2008. V. 167. № 7. P. 847–858. doi: 10.1093/aje/kwm371.
29. Baccarelli A., Mocarelli P., Patterson D.G. Jr., Bonzini M., Pesatory A.C., Caporaso N., Landi M.T. // Environ. Hlth Perspect. 2002. V. 110. № 12. P. 1169–1173.
30. Sweeney M.H., Mocarelli P. // Food Add. Contam. 2000. V. 17. P. 303–316.

Comparison of the Effectiveness of Transepidemal and Intradermal Immunization of Mice with the Vacinia Virus

S. N. Shchelkunov*, A. A. Sergeev, K. A. Titova, S. A. Pyankov, E.V. Starostina, M. B. Borgoyakova, L. A. Kisakova, D. N. Kisakov, L. I. Karpenko, S. N. Yakubitskiy
State Research Center of Virology and Biotechnology VECTOR, Rospotrebnadzor, Koltsovo, Novosibirsk region, 630559 Russia
*E-mail: snshchel@rambler.ru

Received November 10, 2022; in final form, November 30, 2022
DOI: 10.32607/actanaturae.11857

ABSTRACT **The spread of the monkeypox virus infection among humans in many countries outside of Africa, which started in 2022, is now drawing the attention of the medical and scientific communities to the fact that immunization against this infection is sorely needed. According to current guidelines, immunization of people with the first-generation smallpox vaccine based on the vaccinia virus (VACV) LIVP strain, which is licensed in Russia, should be performed via transepidermal inoculation (skin scarification, s.s.). However, the long past experience of using this vaccination technique suggests that it does not ensure virus inoculation into patients' skin with enough reliability. The procedure of intradermal (i.d.) injection of a vaccine can be an alternative to s.s. inoculation. The effectiveness of i.d. vaccination can depend on the virus injection site on the body. Therefore, the aim of this study was to compare the development of the humoral and cellular immune responses in BALB/c mice immunized with the LIVP VACV strain, which was administered either by s.s. inoculation or i.d. injection into the same tail region of the animal. A virus dose of 10^5 pfu was used in both cases. ELISA of serum samples revealed no significant difference in the dynamics and level of production of VACV-specific IgM and IgG after i.d. or s.s. vaccination. A ELISpot analysis of splenocytes from the vaccinated mice showed that i.d. administration of VACV LIVP to mice induces a significantly greater T-cell immune response compared to s.s. inoculation. In order to assess the protective potency, on day 45 post immunization, mice were intranasally infected with lethal doses of either the cowpox virus (CPXV) or the ectromelia virus (ECTV), which is evolutionarily distant from the VACV and CPXV. Both vaccination techniques ensured complete protection of mice against infection with the CPXV. However, when mice were infected with a highly virulent strain of ECTV, 50% survived in the i.d. immunized group, whereas only 17% survived in the s.s. immunized group. It appears, therefore, that i.d. injection of the VACV can elicit a more potent protective immunity against orthopoxviruses compared to the conventional s.s. technique.**

KEYWORDS **orthopoxviruses, vaccinia virus, skin scarification, intradermal injection, antibodies, T cells.**

ABBREVIATIONS **CPXV – cowpox virus; ECTV – ectromelia virus; VACV – vaccinia virus; pfu – plaque forming units; i.d. – intradermal; s.s. – skin scarification; dpi – day post immunization; i.n. – intranasal; LD_{50} – 50% lethal dose of virus.**

INTRODUCTION

During mass vaccination, virus preparations are administered either intramuscularly or subcutaneously, since these techniques are the simplest to perform, ensure accurate vaccine dosage, and do not require a highly qualified staff. However, these body tissues where a vaccine is delivered are immune-poor and usually do not elicit a long-lasting, potent immune response to the administered vaccine [1–3]. Nonetheless, next-generation smallpox vaccines (including the best studied MVA strain) continue to be typically administered intramuscularly or subcutaneously [4, 5].

Skin immunization is a promising alternative to the conventional subcutaneous and intramuscular administration paths. The reason is that not only does the skin act as a physical barrier, preventing penetration of infectious agents into the body, but it also has evolved to become a highly active immune organ. The skin contains various types of dendritic cells, and these professional antigen-presenting cells (APCs) can recognize, assimilate, and process antigens. Importantly, these dendritic cells underpin the necessary association between the innate and adaptive immune responses by migrating into the skin, draining lymph nodes and presenting antigens to T and B cells, thus inducing a pathogen-specific protective immunity. Furthermore, these highly specialized APCs possess significant plasticity, which is modulated by immune signals emanating from other virus-infected skin cells (including keratinocytes, fibroblasts, melanocytes, mast cells, etc.) [1–3, 6].

Transepidermal immunization is historically the first-ever vaccination technique and originates from variolation (variola inoculation). The procedure involved placing infectious material from smallpox patients into skin incisions (skin scarification, s.s.) made in healthy patients. In the late 18th century, E. Jenner proposed inoculating the contents of pustules from people infected with the cowpox rather than infectious material from smallpox patients. This procedure became known as vaccination (vaccine inoculation). Transepidermal immunization was performed using a scalpel, a lancet, or specialized bifurcated needles. Although this vaccination method has made it possible to eradicate smallpox, reliability in delivering viral material into the skin was never sufficiently high [1]. Furthermore, this procedure can be accompanied by the growth of bacterial microflora in the damaged skin [7].

In 1909, C. Mantoux [8] proposed to make intradermal injections using a syringe with a standard needle. This method became actively used in the administration of the BCG anti-tuberculosis vaccine, which was developed in 1921. A century later, the conventional Mantoux technique for intradermal injection is now used only to administer a small number of vaccines. The reason is that this injection method is not easy to perform: the antigen can either be delivered too deep under the skin, or the vaccine may leak out of the injection site [9]. Therefore, staff needs to be specially trained and have experience making such injections.

The recently conducted animal experiments and clinical trials on volunteers have consistently shown that intradermal vaccination elicits a more potent immune response compared to the conventional intramuscular or subcutaneous varieties [10–12]. Furthermore, intradermal vaccination can ensure a robust immune response at a lower vaccine dose [1, 12], which is also important in the case of mass vaccination, when a large number of vaccine doses need to be produced.

Individual studies report the results of experiments on laboratory animals comparing the effectiveness of the immune response against the vaccinia virus (VACV) delivered by different methods: intramuscularly, subcutaneously, intradermally, intraperitoneally, etc. Intradermal injection of VACV has consistently ensured a more robust antiviral immune response compared to other vaccination techniques [10, 13]. The results in these studies also depended on the analyzed VACV strains and virus doses used.

Liu L. et al. [14] demonstrated that inoculating the VACV WR strain highly pathogenic for mice into the scarified tail skin of mice can elicit an immune response stronger than that observed after intradermal injection of this virus into the low back of mice. Skin thickness is known to vary depending on the region of the body [2]. Therefore, the effectiveness of intradermal vaccination can hinge on the virus injection site. All these facts indicate that comparative studies are needed in order to determine how the technique used for inoculating the VACV strain into the skin within the same body area affects the immune response dynamics and level.

The VACV LIVP strain used to design the first-generation smallpox vaccine in Russia [15] was the study object. The study aimed to compare the humoral and T cell-mediated immune responses to vaccination of BALB/c mice with the VACV LIVP strain inoculated into the same tail region by scarification (transepidermally) or by injection with a needle and a syringe using the Mantoux technique (intradermally).

EXPERIMENTAL

Viruses and cells

The clonal variant 14 of the VACV LIVP strain [16], cowpox virus (CPXV) strain GRI-90 [17], and ectromelia virus (ECTV) strain K-1 from the Virus collection and African green monkey kidney cell culture CV-1 from the Cell culture collection of the SRC VB VECTOR were used in this study. The viruses were grown and titrated in the CV-1 cell culture using the procedures described previously [15].

Animals

Female BALB/c mice aged 6–7 weeks (weight, 16–19 g) procured from the husbandry of the SRC VB VECTOR were used for the experiments. The experimental animals were fed a standard diet with ad libitum access to water, in compliance with the veterinary regulations and the guidelines for humane handling and use of animals in research. Animal manipulations were approved by the Bioethics Committee of the SRC VB VECTOR (Protocol No. 01-04.2021 dated April 22, 2021).

Infection of mice

The animals were immunized by intradermal injection (i.d.) or skin scarification (s.s.) using the VACV LIVP at a dose of 10^5 plaque forming units (pfu).

For the i.d. injection, the injection site (the dorsal side of tail, ~ 1 cm from the tail base) was pre-disinfected with 70% ethanol; a needle 30G (0.3 × 13 mm) connected to a syringe was inserted at a small angle, with the needle bevel facing up, to a depth of ~ 2–3 mm under the superficial level of the epidermis. Viral material or saline (control group), 20 μl, was injected slowly, with the expectation that the top skin layers will get delaminated due to the pressure of the fluid (blanching of the skin spreading to both sides of the injection site was indication that the fluid had got into the intradermal space). After the injection, the needle was withdrawn slowly and the injection site was disinfected with 70% ethanol.

For immunization using the s.s. technique, the inoculation site (the dorsal side of the tail, ~ 1 cm from the tail base) was pre-disinfected with 70% ethanol. Once the ethanol had evaporated, 10 skin incisions were made using a needle 26G (0.45 × 16 mm) within the superficial layer of the epidermis. Viral material or saline (5 μl) was immediately placed onto the damaged skin area and was let to be adsorbed by the skin.

Each group consisted of 36 mice.

Sampling of biomaterials from the experimental animals

After the immunization (7, 14, 21, and 28 days post immunization (dpi)) with the VACV, blood samples were collected from the retro-orbital venous sinus of mice (six animals from each group) by puncturing the sinus with a needle 23G (0.6 × 30 mm); the animals were then euthanized by cervical dislocation. Spleens for splenocyte isolation were removed under sterile conditions using forceps and surgical scissors and placed into the transport medium.

Serum specimens were obtained from the individual blood samples of mice by centrifugation of blood cells. Mouse serum specimens were stored at –20°C.

On 42 dpi with VACV, blood samples were collected from the retro-orbital venous sinus intravitally in mice (12 animals from each group) and individual serum specimens were obtained using the procedure described above.

Assessment of the protective potency in immunized mice

On 45 dpi, the groups of virus-immunized and control animals were intranasally (i.n.) infected with CPXV GRI-90 at a dose of 300 LD_{50} (3.2×10^6 pfu) (six animals per group) or with ECTV K-1 at a dose of 300 LD_{50} (7.3×10^3 pfu) (six animals per group). The animals were followed for clinical signs of infection and mortality for 14 days.

The mice were individually weighed every two days. The arithmetic mean body weight of the mice in each group at every time point was calculated and expressed as a percentage of the initial weight. Data were obtained for the group of animals immunized with VACV LIVP, as well as the non-immunized and not-infected group of mice (negative control) and those infected with CPXV GRI-90 or ECTV K-1 (positive control).

Splenocyte isolation

The spleens collected from the immunized mice were mashed onto 70-μm and 40-μm cell strainers (BD Falcon™, Tewksbury, MA, USA). Splenocytes were treated with a red blood cell lysis buffer (ACK Lysis Buffer, Sigma, St. Louis, MO, USA); then, the cells were washed with a completed RPMI 1640 medium and suspended in the completed RPMI 1640 medium with 10% fetal bovine serum, 2 mM L-Gln, and 50 μg/mL gentamycin. The cells were counted with a TC20™ automated cell counter (Bio-Rad, Hercules, CA, USA).

IFN-γ ELISpot assay

The assays were performed using the mouse IFN-γ ELISpot kit (R&D Systems, Inc., Minneapolis, MN, USA) according to the manufacturer's instructions. The splenocytes were plated (100 μL/well) in duplicates 5×10^6 cells/mL and stimulated by a mixture of peptides (corresponding to VACV-specific BALB/c mice H2-d restricted epitopes): SPYAAGYDL, SPGAAGYDL, VGPSNSPTF, KYGRLFNEI, GFIRSLQTI, and KYMWCYSQV [18]. The pooled peptides (100 μL/well) were added at a concentration of 20 μg/mL for each peptide. Non-stimulated and concanavalin A (Con A, 5 μg/mL) stimulated splenocytes were used as the negative and non-specific positive controls, respectively. After an 18-h stimulation period at 37°C in 5% CO_2, the cells were discarded and the plates were incubated for 2 h at 37°C in the presence of anti-IFN-γ detection antibodies.

The plates were washed and the spots were revealed by adding the streptavidin-conjugated alkaline phosphatase and the BCIP/NBT (5-bromo-4-chloro-3′-indolylphosphate/nitro-blue tetrazolium) substrate.

The reaction was stopped by washing the plates with distilled water. The number of IFN-γ-producing cells was counted using an ELISpot reader (Carl Zeiss, Jena, Germany).

Enzyme-linked immunosorbent assay of the serum samples

ELISA of individual mouse serum specimens was performed according to the procedure described earlier [15]. The purified VACV LIVP preparation was used as an antigen. The geometric means of the logarithms of the reciprocal titers of VACV-specific IgM and IgG in the study groups were determined, and the confidence intervals for a 95% confidence level were calculated.

Statistics

The data were analyzed with the GraphPad Prism 9.0 software (GraphPad Software, Inc., San Diego, CA, USA). The results are expressed as a geometric mean with GSD. Data throughout the study were analyzed using repeated-measures two-way ANOVA with the Geisser-Greenhouse correction. Multiple comparisons were performed using a Tukey test. The statistical analysis was conducted at a 95% confidence level. A *P* value less than 0.05 was considered statistically significant.

RESULTS

Intradermal injection of VACV LIVP to mice induces a stronger cell-mediated immune response compared to virus inoculation by skin scarification

Changes in the T-cell immune response in LIVP-vaccinated BALB/c mice over time were investigated using the IFN-γ ELISpot technique. The mice were split into several groups (six animals per group). The animals were inoculated with the VACV LIVP either i.d. (1 cm from the tail base) or s.s. (1 cm from the tail base) at a dose of 10^5 pfu/animal. The spleens for performing ELISpot assay were removed individually from six animals in each study group on 7, 14, 21, and 28 dpi. Intact (non-immunized) mice were used as control.

The intensity of the T cell-mediated immune response in the immunized mice was determined according to the number of splenocytes producing IFN-γ in response to the stimulation with peptides from the immunodominant VACV proteins [19]. The results shown in *Fig. 1* demonstrate that a potent VACV-specific T cell-mediated immune response was elicited in all immunized mice. Meanwhile, the splenocytes in the control animals did not produce IFN-γ.

After s.s. inoculation of VACV LIVP, on 7 dpi only a low level of VACV-specific T cell-mediated immunity was induced in mice, reaching its maximum on 14 dpi and declining significantly on 21 and 28 dpi (*Fig. 1*).

After i.d. injection, an intensive T cell-mediated immune response developed in mice as early as on 7 dpi, slightly increased by 14 dpi, and remained high during the entire follow-up period (up to 28 dpi).

On days 7, 21, and 28, the level of T cell response in i.d. vaccinated mice significantly exceeded that in the groups of mice s.s. inoculated with VACV LIVP (*Fig. 1*).

No difference in the dynamics of developing humoral immunity in mice in response to inoculation of VACV LIVP by intradermal injection or skin scarification was revealed

Individual blood samples were collected from the retro-orbital venous sinus in mice on 7, 14, 21, 28, and 42 dpi to obtain serum specimens, which were then analyzed by ELISA; the preparation of VACV LIVP virions was used as an antigen.

Serum samples from six animals were analyzed at each time point in each group. The geometric means of the logarithms of reciprocal titers of VACV-specific IgM and IgG were calculated. The maximum level of VACV-specific IgM was observed in mice on 21 dpi (*Fig. 2*), while the maximum level of VACV-specific IgG production was observed on 28 dpi (*Fig. 3*).

No statistically significant differences in the IgM or IgG levels in serum samples were revealed between the groups of mice immunized by i.d. injection and s.s. inoculation of the VACV LIVP strain (*Figs. 2,3*).

Intradermal injection of VACV LIVP to mice provides greater protective potency than inoculation of this virus by skin scarification

In order to understand how the levels of humoral and cell-mediated immunity developing in response to the immunization of the mice with the VACV LIVP affect their protective potency against a challenge with a lethal orthopoxvirus infection, the mice were i.n. infected with lethal doses of CPXV GRI-90 (six animals per group) or ECTV K-1 (six animals per group) on day 45 post i.d. or s.s. inoculation of the VACV LIVP. The mice were followed up for 14 days; clinical manifestations of the infection and death of the animals were documented. Every two days, mice were weighed to determine the dynamics of body weight change.

After the mice had been infected i.n. with CPXV at a dose of 3.2 × 10^6 pfu (300 LD_{50}), the animals in the study groups started displaying signs of disease and their body weight declined transiently on days 4–8 without statistically significant differences (*Fig. 4A*). All the animals in the positive control group had died

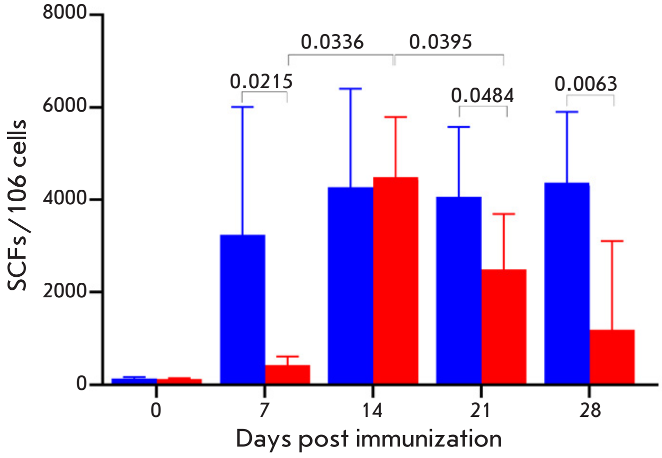


Fig. 1. Assessment of T cell-mediated immunity in BALB/c mice immunized with VACV LIVP (six mice per group) by IFN-γ ELISpot assay. Splenocytes were stimulated with a pool of virus-specific peptides during 24 h. Blue bars – i.d. injection of the VACV LIVP; red bars – s.s. inoculation of the VACV LIVP. The diagrams show the geometric mean with GSD. The Y axis shows the number of spots (the number of IFNγ-producing cells) per 10^6 splenocytes. Day 0 – the level of T cell-mediated immune response for non-immunized mice. The statistical analysis was performed using the GraphPad Prism 9.0 software. *P* values are above horizontal brackets

by day 6, while all the mice in study groups had recovered (*Fig. 4B*).

After the mice had been infected i.n. with the highly pathogenic ECTV at a dose of 7.3 × 10^3 pfu (300 LD_{50}), signs of disease were observed in study groups on days 6–10 and the animals' body weights declined transiently without statistically significant differences (*Fig. 5A*). All the animals in the positive control groups had died by day 8. Half of the mice in the group of animals vaccinated by i.d. injection survived, while only 17% of the animals vaccinated by s.s. inoculation of the virus survived (*Fig. 5B*).

DISCUSSION

The skin possesses properties that make it an excellent site for vaccination. It is an immune-rich organ and contains components that efficiently induce both humoral and cell-mediated immunity in response to infection/vaccination [1–3]. There are two techniques for cutaneous vaccination: the historically older method of transepidermal inoculation or skin scarification (s.s.) and the technique of intradermal injection (i.d.), which was proposed in the early 20th century [8]. Each of these methods has advantages and shortcomings.

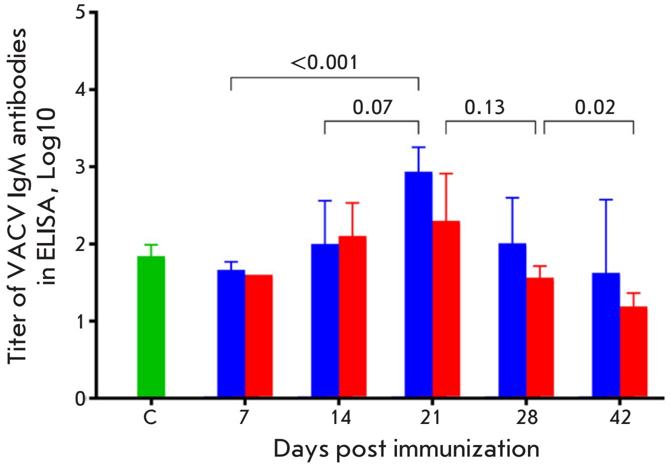


Fig. 2. Concentration of VACV-specific IgM in the serum samples of mice immunized with VACV LIVP at a dose of 10^5 pfu determined by ELISA. Blue bars – i.d. injection of the VACV LIVP; red bars – s.s. inoculation of the VACV LIVP. C (control) – serum samples from mice that received saline. The diagrams show the geometric mean with GSD. The statistical analysis was performed using the GraphPad Prism 9.0 software. *P* values are above horizontal brackets

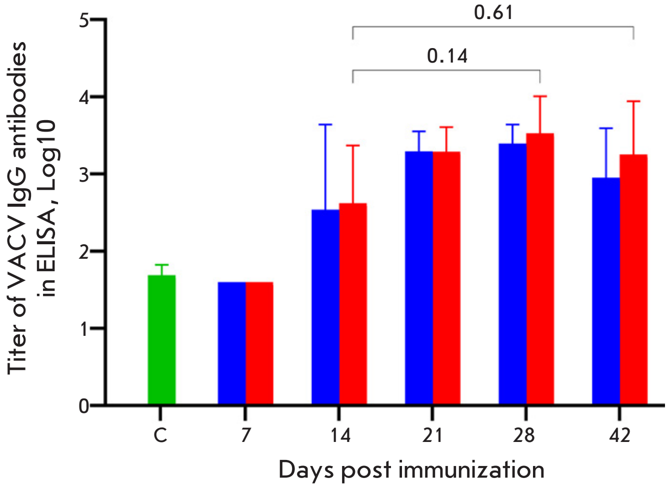


Fig. 3. Concentration of VACV-specific IgG in the serum samples of mice immunized with the VACV LIVP at a dose of 10^5 pfu determined by ELISA. Blue bars – i.d. injection of the VACV LIVP; red bars – s.s. inoculation of the VACV LIVP. C (control) – serum samples from mice that received saline. The diagrams show the geometric mean with GSD. The statistical analysis was performed using the GraphPad Prism 9.0 software. *P* values are above horizontal brackets

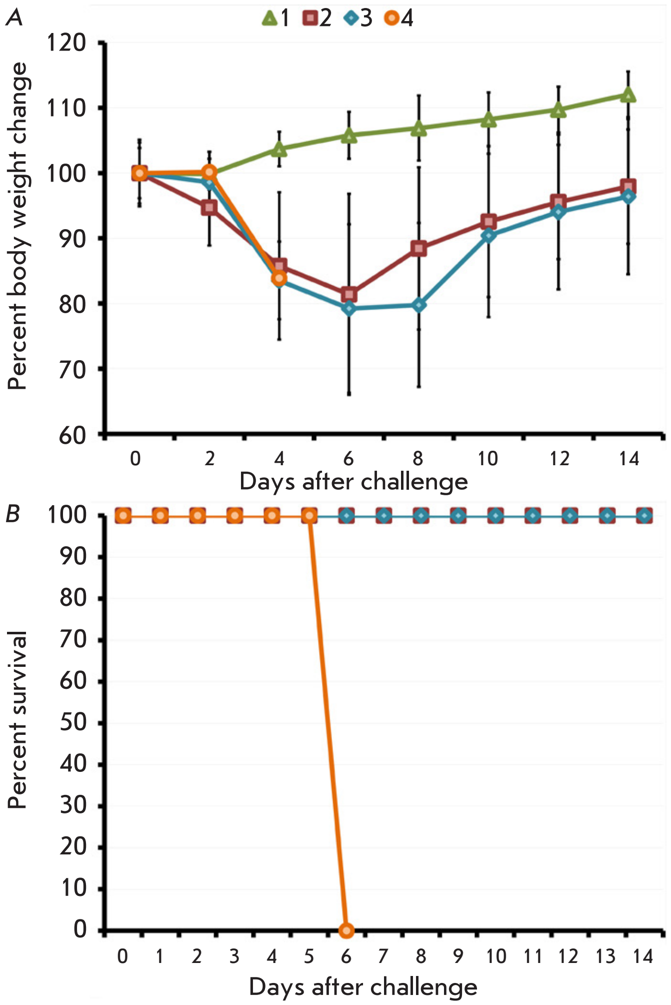


Fig. 4. The dynamics of body weight change (*A*) and death of mice (*B*) immunized with VACV LIVP at a dose of 10^5 pfu after i.n. infection with CPXV GRI-90 at a dose of 300 LD_{50}. Data for groups consisting of six animals immunized using the s.s. (2) or i.d. (3) technique, as well as the groups consisting of non-immunized animals either non-infected (1) or infected with CPXV GRI-90 (4), are shown

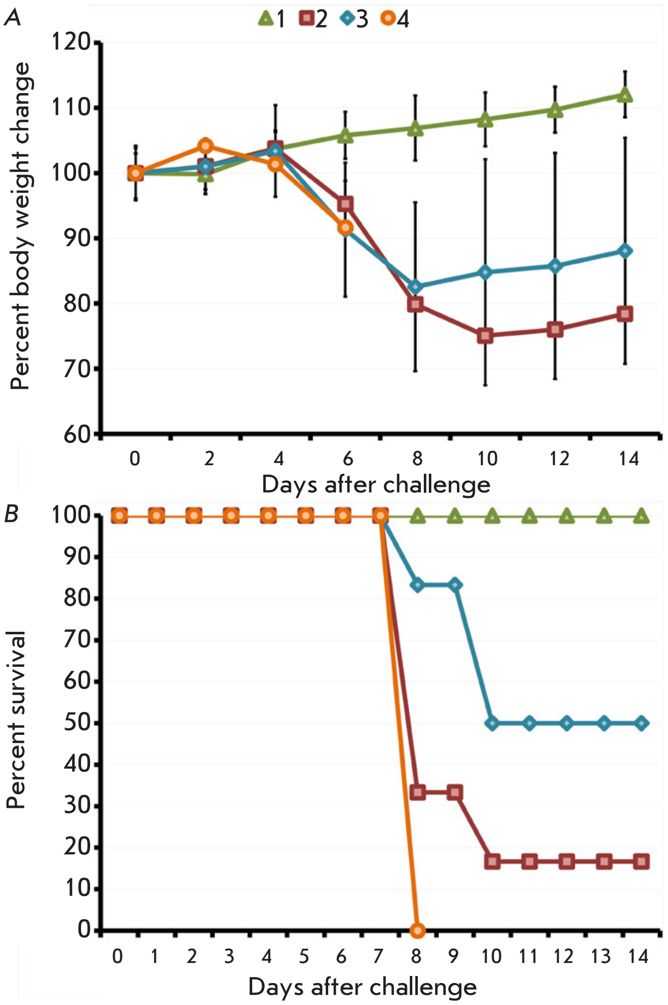


Fig. 5. The dynamics of body weight change (*A*) and death of mice (*B*) immunized with VACV LIVP at a dose of 10^5 pfu after i.n. infection with ECTV K-1 at a dose of 300 LD_{50}. Data for groups consisting of six animals immunized using the s.s. (2) or i.d. (3) technique, as well as the groups consisting of non-immunized animals either non-infected (1) or infected with ECTV K-1 (4), are shown

The technique of s.s. inoculation is relatively simple, but the skin cover is disrupted when used in that way. Thus, a local inflammatory response is induced and it is difficult to ensure dosage accuracy. The i.d. injection using a needle and a syringe causes minimal skin damage and allows one to dose the vaccine and inject it into the target skin layer more accurately.

Despite the long history of using both the s.s. and i.d. vaccination techniques, no fully correct comparison of the immunogenic and protective effectiveness of these two methods upon inoculation of the VACV in animal models has been performed. Such a conclusion can be drawn because in most studies comparing the s.s. and i.d. techniques, the VACV was inoculated into different body sites of laboratory mice [19]. The results of our preliminary experiments have shown that the body site of mice into which the virus preparation is inoculated significantly affects the immune response level upon i.d. injection of the VACV. In order to eliminate this effect, we have compared the s.s. and i.d. techniques when the same dose of the VACV is inoculated into the same site at the mouse tail.

BALB/c mice and the VACV LIVP strain were used as study objects. The VACV LIVP at a dose of 10^5 pfu was inoculated either i.d. or s.s. to mice into the tail skin (1 cm from the tail base). For each of the two studied vaccination methods, blood was sampled from the retro-orbital venous sinus in six animals at each time point (7, 14, 21, and 28 dpi) and individual serum samples for the analysis of the levels of VACV-specific antibodies were obtained. Next, spleens were removed from each animal to isolate splenocytes and perform a IFN-γ ELISpot assay. Intact (non-immunized) mice were used as control.

The intensity of the T cell-mediated immune response in immunized mice was determined according to the number of splenocytes producing IFN-γ in response to stimulation with peptides from the immunodominant VACV proteins (*Fig. 1*). Only a low level of VACV-specific T cell-mediated response was induced by s.s. inoculation of the VACV LIVP on 7 dpi; it reached its maximum on 14 dpi and began declining significantly by 21 and 28 dpi. After i.d. injection, an intensive T cell-mediated immune response developed in mice as early as on 7 dpi and remained so during the entire follow-up period (up to 28 dpi). On 7, 21, and 28 dpi, the level of the T cell response in i.d.-vaccinated mice significantly exceeded that in the groups of mice s.s. inoculated with VACV LIVP (*Fig. 1*). Hence, i.d. immunization with the VACV LIVP induces a more potent and lansting T cell-mediated immune response in mice compared to s.s. vaccination.

In the remaining mice in the study and control groups (12 animals per group), blood was sampled intravitally from the retro-orbital venous sinus on 42 dpi, and individual serum samples were obtained. ELISA of all the serum samples of the immunized mice revealed no statistically significant difference in the dynamics and level of production of VACV-specific IgM (*Fig. 2*) and IgG (*Fig. 3*) after both the i.d. and s.s. vaccinations. The maximum IgM and IgG levels were observed on 21 and 28 dpi, respectively.

In order to assess the protective immunity that developed as a result of i.d. or s.s. vaccination, six mice per group were infected i.n. with highly lethal doses of CPXV GRI-90 or ECTV K-1 on 45 dpi. Both vaccination methods were found to completely protect mice against infection with CPXV at a dose of 300 LD_{50} (*Fig. 4*). However, the vaccinated animals had only partial protection after being i.n. infected with a highly virulent ECTV (300 LD_{50}), which is relatively evolutionarily distant from the VACV and CPXV [20] (*Fig. 5*). Meanwhile, 50% of the mice immunized by i.d. injection survived; the percentage of surviving mice immunized by s.s. inoculation was 17%.

These findings allow us to infer that, although humoral immunity makes the greatest contribution to the protection against a challenge with the orthopoxvirus infection [21–23], the level of cell-mediated immunity that develops in response to vaccination is also important. A conclusion can also be drawn that intradermal injection of the VACV can ensure a more potent protective immunity compared to the conventional skin scarification technique because of the stronger T cell-mediated response.

The results obtained in this study differ from the findings published earlier by T.S. Kupper et al. [14, 19], who revealed that the VACV exhibits a higher immunogenicity and protectivity upon s.s. immunization of mice compared to the i.d. and other routes of injection of the virus. In those studies, C57BL/6 mice were immunized with the non-replicating VACV MVA strain and protectivity against a lethal respiratory challenge with the VACV WR strain was assessed. For different routes of administration of the viruses, different body parts of mice were challenged.

A different, BALB/c, line of mice was used in our study, and the animals were immunized with the replicating VACV LIVP strain. The protectivity of the immunized mice against a lethal respiratory challenge with the heterologous orthopoxviruses CPXV and ECTV was assessed. Preliminary experiments have revealed that the immunogenicity of the VACV LIVP strain differs significantly upon i.d. injection of the virus into different body sites of mice. Therefore, the VACV LIVP strain was injected into the same region of mouse tail skin in order to properly compare the efficacies of the s.s. and i.d. routes of immunization. This fact seems to be responsible for the discrepancies between our results and the data published previously [14, 19].

The advances in modern techniques of intradermal injection of vaccines will simplify this promising approach to antiviral immunization and increase its reliability [1–3, 24]. ●

This work was supported by the Russian Science Foundation (grant No. 19-14-00006-P).

REFERENCES

1. Kim Y.C., Jarrahian C., Zehrung D., Mitragotri S., Prausnitz M.R. // Curr. Topics Microbiol. Immunol. 2012. V. 351. P. 77–112.
2. Gamazo C., Pastor Y., Larraneta E., Berzosa M., Irache J.M., Donnelly R.F. // Ther. Deliv. 2019. V. 10. P. 63–80.
3. Hettinga J., Carlisle R. // Vaccines. 2020. V. 8. P. 534.
4. Vollmar J., Arndtz N., Eckl K.M., Thomsen T., Petzold B., Mateo L., Schlereth B., Handley A., King L., Hulsemann V., et al. // Vaccine. 2006. V. 24. P. 2065–2070.
5. Jackson L.A., Frey S.E., El Sahly H.M., Mulligan M.J., Winokur P.L., Kotloff K.L., Campbell J.D., Atmar R.L.,

Graham I., Anderson E.J., et al. // Vaccine. 2017. V. 35. P. 1675–1682.
6. Lei V., Petty A.J., Atwater A.R., Wolfe S.A., MacLeod A.S. // Front. Immunol. 2020. V. 11. 593901.
7. Shmeleva E.V., Gomez de Aguero M., Wagner J., Enright A.J., Macpherson A.J., Ferguson B.J., Smith G.L. // PLoS Pathog. 2022. V. 18. e1009854.
8. Mantoux C. // C. R. Hebd. Seanc. Acad. Sci., Paris. 1909. V. 148. P. 996–998.
9. Tarnow K., King N. // Appl. Nursing Res. 2004. V. 17. P. 275–282.
10. Egunsola O., Clement F., Taplin J., Mastikhina L., Li J.W., Lorenzetti D.L., Dowsett L.E., Noseworthy T. // JAMA Network Open. 2021. V. 4. P. e2035693.
11. Wilck M.B., Seaman M.S., Baden L.R., Walsh S.R., Grandpre L.E., Devoy C., Giri A., Kleinjan J.A., Noble L.C., Stevenson K.E., et al. // J. Infect. Dis. 2010. V. 201. P. 1361–1370.
12. Schnyder J.L., De Pijper C.A., Garcia Garrido H.M., Daams J.G., Goorhuis A., Stijnis C., Schaumburg F., Grobusch M.P. // Trav. Med. Infect. Dis. 2020. V. 37. 101868.
13. Hughes L.J., Townsend M.B., Gallardo-Romero N., Hutson C.L., Patel N., Dotty J.B., Salzer J.S., Damon I.K., Carroll D.S., Satheshkumar P.S., et al. // Virology. 2020. V. 544. P. 55–63.
14. Liu L., Zhong Q., Tian T., Dubin K., Athale S.K., Kupper T.S. // Nat. Med. 2010. V. 16. P. 224–227.
15. Shchelkunov S.N., Yakubitskiy S.N., Sergeev A.A., Kabanov A.S., Bauer T.V., Bulichev L.E., Pyankov S.A. // Viruses. 2020. V. 12. P. 795.
16. Yakubitskiy S.N., Kolosova I.V., Maksyutov R.A., Shchelkunov S.N. // Acta Naturae. 2015. № 4 (27). V. 7. P. 113–121.
17. Shchelkunov S.N., Safronov P.F., Totmenin A.V., Petrov N.A., Ryazankina O.I., Gutorov V.V., Kotwal G.J. // Virology. 1998. V. 243. P. 432–460.
18. Shchelkunov S.N., Sergeev A.A., Yakubitskiy S.N., Titova K.A., Pyankov S.A., Kolosova I.V., Starostina E.V., Borgoyakova M.B., Zadorozhny A.M., Kisakov D.N., et al. // Viruses. 2021. V. 13. P. 1631.
19. Pan Y., Liu L., Tian T., Zhao J., Park C.O., Lofftus S.Y., Stingley C.A., Yan Y., Mei S., Liu X., et al. // NPJ Vaccines. 2021. V. 6. P. 1.
20. Carroll D.S., Emerson G.L., Li Y., Sammons S., Olson V., Frace M., Nakazawa Y., Czerny C.P., Tryland M., Kolodziejek J., et al. // PLoS One. 2011. V. 6. P. e23086.
21. Belyakov I.M., Earl P., Dzutsev A., Kuznetsov V.A., Lemon M., Wyatt L.S., Snyder J.T., Ahlers J.D., Franchini G., Moss B., Berzofsky J.A. // Proc. Natl. Acad. Sci. USA. 2003. V. 100. P. 9458–9463.
22. Moss B. // Immunol. Rev. 2011. V. 239. P. 8–26. doi:10.1111/j.1600-065X.2010.00975.x
23. Shchelkunov S.N., Shchelkunova G.A. // Acta Naturae. 2020. V. 12. № 1 (44). P. 33–41.
24. Lambert P.H., Laurent P.E. // Vaccine. 2008. V. 26. P. 3197–3208.

GENERAL RULES

Acta Naturae publishes experimental articles and reviews, as well as articles on topical issues, short reviews, and reports on the subjects of basic and applied life sciences and biotechnology.

The journal *Acta Naturae* is on the list of the leading periodicals of the Higher Attestation Commission of the Russian Ministry of Education and Science. The journal Acta Naturae is indexed in PubMed, Web of Science, Scopus and RCSI databases.

The editors of *Acta Naturae* ask of the authors that they follow certain guidelines listed below. Articles which fail to conform to these guidelines will be rejected without review. The editors will not consider articles whose results have already been published or are being considered by other publications.

The maximum length of a review, together with tables and references, cannot exceed 50,000 characters with spaces (approximately 30 pages, A4 format, 1.5 spacing, Times New Roman font, size 12) and cannot contain more than 16 figures.

Experimental articles should not exceed 30,000 symbols (approximately 15 pages in A4 format, including tables and references). They should contain no more than ten figures.

A short report must include the study's rationale, experimental material, and conclusions. A short report should not exceed 12,000 symbols (5–6 pages in A4 format including no more than 12 references). It should contain no more than three figures.

The manuscript and all necessary files should be uploaded to www.actanaturae.ru:
1) text in Word 2003 for Windows format;
2) the figures in TIFF format;
3) the text of the article and figures in one pdf file;
4) the article's title, the names and initials of the authors, the full name of the organizations, the abstract, keywords, abbreviations, figure captions, and Russian references should be translated to English;
5) the cover letter stating that the submitted manuscript has not been published elsewhere and is not under consideration for publication;
6) the license agreement (the agreement form can be downloaded from the website www.actanaturae.ru).

MANUSCRIPT FORMATTING

The manuscript should be formatted in the following manner:

- Article title. Bold font. The title should not be too long or too short and must be informative. The title should not exceed 100 characters. It should reflect the major result, the essence, and uniqueness of the work, names and initials of the authors.
- The corresponding author, who will also be working with the proofs, should be marked with a footnote *.
- Full name of the scientific organization and its departmental affiliation. If there are two or more scientific organizations involved, they should be linked by digital superscripts with the authors' names. Abstract. The structure of the abstract should be very clear and must reflect the following: it should introduce the reader to the main issue and describe the experimental approach, the possibility of practical use, and the possibility of further research in the field. The average length of an abstract is 20 lines (1,500 characters).
- Keywords (3 – 6). These should include the field of research, methods, experimental subject, and the specifics of the work. List of abbreviations.
- INTRODUCTION
- EXPERIMENTAL PROCEDURES
- RESULTS AND DISCUSSION
- CONCLUSION

  The organizations that funded the work should be listed at the end of this section with grant numbers in parenthesis.
- REFERENCES

  The in-text references should be in brackets, such as [1].

RECOMMENDATIONS ON THE TYPING AND FORMATTING OF THE TEXT

- We recommend the use of Microsoft Word 2003 for Windows text editing software.
- The Times New Roman font should be used. Standard font size is 12.
- The space between the lines is 1.5.
- Using more than one whole space between words is not recommended.
- We do not accept articles with automatic referencing; automatic word hyphenation; or automatic prohibition of hyphenation, listing, automatic indentation, etc.
- We recommend that tables be created using Word software options (Table → Insert Table) or MS Excel. Tables that were created manually (using lots of spaces without boxes) cannot be accepted.
- Initials and last names should always be separated by a whole space; for example, A. A. Ivanov.
- Throughout the text, all dates should appear in the "day.month.year" format, for example 02.05.1991, 26.12.1874, etc.
- There should be no periods after the title of the article, the authors' names, headings and subheadings, figure captions, units (s – second, g – gram, min – minute, h – hour, d – day, deg – degree).
- Periods should be used after footnotes (including those in tables), table comments, abstracts, and abbreviations (mon. – months, y. – years, m. temp. – melting temperature); however, they should not be used in subscripted indexes (T_m – melting temperature; $T_{p.t}$ – temperature of phase transition). One exception is mln – million, which should be used without a period.
- Decimal numbers should always contain a period and not a comma (0.25 and not 0,25).
- The hyphen ("-") is surrounded by two whole spaces, while the "minus," "interval," or "chemical bond" symbols do not require a space.
- The only symbol used for multiplication is "×"; the "×" symbol can only be used if it has a number to its

right. The "·" symbol is used for denoting complex compounds in chemical formulas and also noncovalent complexes (such as DNA·RNA, etc.).

- Formulas must use the letter of the Latin and Greek alphabets.
- Latin genera and species' names should be in italics, while the taxa of higher orders should be in regular font.
- Gene names (except for yeast genes) should be italicized, while names of proteins should be in regular font.
- Names of nucleotides (A, T, G, C, U), amino acids (Arg, Ile, Val, etc.), and phosphonucleotides (ATP, AMP, etc.) should be written with Latin letters in regular font.
- Numeration of bases in nucleic acids and amino acid residues should not be hyphenated (T34, Ala89).
- When choosing units of measurement, SI units are to be used.
- Molecular mass should be in Daltons (Da, KDa, MDa).
- The number of nucleotide pairs should be abbreviated (bp, kbp).
- The number of amino acids should be abbreviated to aa.
- Biochemical terms, such as the names of enzymes, should conform to IUPAC standards.
- The number of term and name abbreviations in the text should be kept to a minimum.
- Repeating the same data in the text, tables, and graphs is not allowed.

GUIDENESS FOR ILLUSTRATIONS

- Figures should be supplied in separate files. Only TIFF is accepted.
- Figures should have a resolution of no less than 300 dpi for color and half-tone images and no less than 600 dpi.
- Files should not have any additional layers.

REVIEW AND PREPARATION OF THE MANUSCRIPT FOR PRINT AND PUBLICATION

Articles are published on a first-come, first-served basis. The members of the editorial board have the right to recommend the expedited publishing of articles which are deemed to be a priority and have received good reviews.

Articles which have been received by the editorial board are assessed by the board members and then sent for external review, if needed. The choice of reviewers is up to the editorial board. The manuscript is sent on to reviewers who are experts in this field of research, and the editorial board makes its decisions based on the reviews of these experts. The article may be accepted as is, sent back for improvements, or rejected.

The editorial board can decide to reject an article if it does not conform to the guidelines set above.

The return of an article to the authors for improvement does not mean that the article has been accepted for publication. After the revised text has been received, a decision is made by the editorial board. The author must return the improved text, together with the responses to all comments. The date of acceptance is the day on which the final version of the article was received by the publisher.

A revised manuscript must be sent back to the publisher a week after the authors have received the comments; if not, the article is considered a resubmission.

E-mail is used at all the stages of communication between the author, editors, publishers, and reviewers, so it is of vital importance that the authors monitor the address that they list in the article and inform the publisher of any changes in due time.

After the layout for the relevant issue of the journal is ready, the publisher sends out PDF files to the authors for a final review.

Changes other than simple corrections in the text, figures, or tables are not allowed at the final review stage. If this is necessary, the issue is resolved by the editorial board.

FORMAT OF REFERENCES

The journal uses a numeric reference system, which means that references are denoted as numbers in the text (in brackets) which refer to the number in the reference list.

For books: the last name and initials of the author, full title of the book, location of publisher, publisher, year in which the work was published, and the volume or issue and the number of pages in the book.

For periodicals: the last name and initials of the author, title of the journal, year in which the work was published, volume, issue, first and last page of the article. Must specify the name of the first 10 authors. Ross M.T., Grafham D.V., Coffey A.J., Scherer S., McLay K., Muzny D., Platzer M., Howell G.R., Burrows C., Bird C.P., et al. // Nature. 2005. V. 434. № 7031. P. 325–337.

References to books which have Russian translations should be accompanied with references to the original material listing the required data.

References to doctoral thesis abstracts must include the last name and initials of the author, the title of the thesis, the location in which the work was performed, and the year of completion.

References to patents must include the last names and initials of the authors, the type of the patent document (the author's rights or patent), the patent number, the name of the country that issued the document, the international invention classification index, and the year of patent issue.

The list of references should be on a separate page. The tables should be on a separate page, and figure captions should also be on a separate page.

The following e-mail addresses can be used to contact the editorial staff: actanaturae@gmail.com, tel.: (495) 727-38-60.