OXFORD

Bioinformatics in Russia: history and present-day landscape

Muhammad A. Nawaz (iD)[1,2,*], Igor E. Pamirsky[1,3], Kirill S. Golokhvast[1,3]

[1]Advanced Engineering School (Agrobiotek), National Research Tomsk State University, Lenin Ave, 36, Tomsk Oblast, Tomsk 634050, Russia
[2]Centre for Research in the Field of Materials and Technologies, National Research Tomsk State University, Lenin Ave, 36, Tomsk Oblast, Tomsk 634050, Russia
[3]Siberian Federal Scientific Centre of Agrobiotechnology, Centralnaya st., 2b, Presidium, Krasnoobsk, 633501, Novosibirsk Oblast, Russia

*Corresponding author. Muhammad A. Nawaz, Advanced Engineering School (Agrobiotek), National Research Tomsk State University, Lenin Ave, 36, Tomsk Oblast, Tomsk 634050, Russia. E-mail: amjad_ucauos@yahoo.com

Abstract

Bioinformatics has become an interdisciplinary subject due to its universal role in molecular biology research. The current status of Russia's bioinformatics research in Russia is not known. Here, we review the history of bioinformatics in Russia, present the current landscape, and highlight future directions and challenges. Bioinformatics research in Russia is driven by four major industries: information technology, pharmaceuticals, biotechnology, and agriculture. Over the past three decades, despite a delayed start, the field has gained momentum, especially in protein and nucleic acid research. Dedicated and shared centers for genomics, proteomics, and bioinformatics are active in different regions of Russia. Present-day bioinformatics in Russia is characterized by research issues related to genetics, metagenomics, OMICs, medical informatics, computational biology, environmental informatics, and structural bioinformatics. Notable developments are in the fields of software (tools, algorithms, and pipelines), use of high computation power (e.g. by the Siberian Supercomputer Center), and large-scale sequencing projects (the sequencing of 100 000 human genomes). Government funding is increasing, policies are being changed, and a National Genomic Information Database is being established. An increased focus on eukaryotic genome sequencing, the development of a common place for developers and researchers to share tools and data, and the use of biological modeling, machine learning, and biostatistics are key areas for future focus. Universities and research institutes have started to implement bioinformatics modules. A critical mass of bioinformaticians is essential to catch up with the global pace in the discipline.

Keywords: bioinformatics; computational biology; Russia; omics; genomics; databases

Introduction

Bioinformatics is an interdisciplinary field that uses the concepts and techniques from applied biology, mathematics, statistics, data science, computer science, information technology, physics, and biochemistry [1]. It is also considered a switchboard between these sciences because of its characteristic of ease of data translation from one discipline to another [2]. Early use of the term "bioinformatics" started in the last decade of the twentieth century, when sequencing projects, e.g. human genome sequence [3], began to produce large amounts of data and the need erose for databases to host the data and tools to analyze it [4]. In general, bioinformatics integrates methods and models to collect, classify, store, and analyze biological data generated from molecular genetics, genomics, proteomics, phenomics, and metabolomics-related experiments. Early bioinformatics began with the sequencing of insulin, which later led to the development of protein sequencing methods, analyses, and the use of computational methods [5]. Advances in bioinformatics (since 1950s) are due to parallel developments in deoxyribonucleic acid (DNA) sequencing methods and computer science, cost reductions in whole genome sequencing (WGS) [6], improvements in massively parallel sequencing methods, and extreme competition among key market players (https://3billion. io/index; accessed on 21 December 2023). Previously defined as a "megascience" by the Organization of Economic Co-operation and Development, it has now become an interdisciplinary subject primarily because of its integration into medicine and health care, and to a greater extent, into agriculture. It has been estimated that by 2025, the annual acquisition and storage of genomics data will amount to 1 Zetta bases and 2 to 40 exa bases per year, respectively, making genomics the most demanding scientific field in terms of data acquisition, storage, sharing, and analysis [7]. In this article we present the history and current state of research and development in the field of bioinformatics in Russia.

Landscape of bioinformatics in Russia
State of the key industries that drive developments in bioinformatics

Before looking at the state of bioinformatics in Russia, it is important to understand the state of the key industries that drive the developments in bioinformatics. Four major industries, i.e. information technology, pharmaceuticals, biotechnology, and agriculture can be considered as the drivers of the bioinformatics field.

Russia's aggregate score of 88.9 in the information and communication technologies (ICT) Development Index in 2021 [8],

indicate that its IT market (12% per year) is growing faster than the global market (5%) in dollar terms. By 2019, the gross value added of the Russian IT industry reached 945 billion rubles. The Russian ICT market is projected to reach a value of US$ 51.47 billion by 2027, up from US$ 34.26 billion in 2022. The cumulative revenue opportunity for ICT providers in Russia is estimated at US$ 252.08 billion between 2022 and 2027 [9]. Similarly, the Russian IT market is projected to grow to RUB 7 trillion by 2030 (Fig. 1). This growth in the IT industry has a direct impact on bioinformatics and quantum computational biology [12] due to the components of high-performance computing (HPC) i.e. hardware, operating system, and networks [13]. Notably, the availability of supercomputers has driven development of algorithms for biological problem-solving. The Center for Collective Use "Siberian Super Computer Center (SSCC)" [14] allows solving biological problems related to omics, biological process modeling, e.g. modeling of molecular genetic systems MGSmodeller, evolution, molecular dynamics, mathematical problems of bioinformatics, and genome sequence analysis, e.g. ICGenomics [15]. The HPC has allowed bioinformaticians to develop new software for genome sequencing data (Chip-seq, ChIA-PET, and Hi-C) [15]. Hardware acceleration is also linked with developments in genomics data analysis as it solves several challenges, such as read alignment and mapping. This is usually associated with the use of HPC clusters, cloud computing clusters, graphic processing units, field programmable gate arrays, etc. [16]. Researchers from the Laboratory of Statistical Multiomics and Bioinformatics, Ufa Federal Research Center of the Russian Academy of Sciences (RAS) are developing an online library of analytical tools (ANTE-OMICS platform) and an electronic database for clinical data [17]. However, as in other countries, the use of principles and practices of software engineering for the development of bioinformatics tools is still limited [18]. Moreover, these developments seem to be directly related to the IT industry, but the components of HPC used in Russian bioinformatics industry are imported. Thus, in a way, the Russian bioinformatics industry relies on IT imports. This was recently realized when the government introduced anti-Russian sanctions move, such as the approval of the plan for the development of electronics and microelectronics in Russia by 2030 [19].

As of 2024, Russia's pharmaceutical market values 2805 billion Rubles (Fig. 1), while it ranked 32nd in international market (in 2022) with an increasing import and export value of 13 billion and 705 million USD, respectively [20]. Pharmaceutical production is one of the priority industries of the country and is expected to grow to a volume of 3757 billion Rubles by 2030 (Fig. 1). This growth is driven by escalating healthcare expenditures, improved lifestyles, and a ballooning aged population. The need for drug discovery, metabolism, target identification, compound optimization, and prediction of drug-target interaction leverages the development of pipelines, algorithms, software, and databases [21]. A notable example in this regard is how the onset of SARS-CoV-2 drove the development of vaccines, detection tests, and a search for potential inhibitors (Information Box-1). Pharmaceutical industry is also shifting the landscape of the bioinformatics where the integration of WGS into clinical practice is becoming a routine, e.g. panels for immune disorders, hereditary disorders/diseases, cancer, and exome sequencing (https://price.genomed.ru/; accessed on 01 August 2024).

Information Box-1

Bioinformatics proved to be a key ally in the fight against the COVID-19 pandemic [162]. The first confirmed

---

COVID-19 case was registered in Russia in late January 2020. At the same time, the first SARS-CoV-2 genome was released publicly on the Global Initiative on Sharing Avian Influenza Data [163], which initiated a global race to develop vaccines and search for inhibitors. By 18 March 2020, Russia had reported SARS-CoV-2 genome isolated from a Russian patient by Smorodintsev Research Institute of Influenza, the WHO National Influenza Centre of Russia [164]. The center also established Russian Consortium for Sequencing Coronavirus Genomes. Several Russian institutes joined the fight against the pandemic. These include FBIS Central Research Institute of Epidemiology of Rospotrebnadzor, Vector State Research Centre of Virology and Biotechnology, Gamaleya Research Institute of Epidemiology and Microbiology, Chumakov Federal Scientific Centre for Research and Development of Immune-and-Biological Products, Petrovax, Human Stem Cell Institute, St. Petersburg Research Institute of Vaccines and Sera, and Biocad, etc. The availability of the SARS-CoV-2 genome triggered the development of vaccines and the first COVID-19 vaccine Sputnik-V was registered on 11 August 2020 [165, 166]. Other laboratories also initiated trials and registered several vaccine candidates (protein subunit, non-replicating viral vector, and inactivated type) and registered them such as rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine [167], EpiVac-Corona [168], Sputnik Light [169], Aurora-CoV, CoviVac [170], Gam-COVID-Vac [165, 166], and KoviVac [171, 172]. These vaccines had a notable impact on the global response to the SARS-CoV-2 pandemic during 2020 to 2022. By 2022, >90 000 records of viruses were available in the Russian coronavirus genome database (VGARus; established in June 2021). After vaccination, it is crucial to analyze the recognition and possible hydrolysis of individual oligopeptides corresponding to linear epitopes of viral antigens to predict antigen efficacy for vaccination. In this regard, a study reported the catalytic activity of antibodies generated after COVID-19 and Sputnik V adenovirus vaccination against nine oligopeptides corresponding to the SARS-CoV-2 S protein [173]. Another research group reported the identification of antibody-mediated hydrolysis sites of oligopeptides corresponding to the SARS-CoV-2 S protein [174]. In collaboration with University of Ottwa, Krasnoyarsk Science Centre of SB RAS discovered DNA aptamers targeting SARS-CoV-2 nucleocapsid protein and protein binding epitopes [175]. Even after these reports, research on the inhibitors of SARS-CoV-2 viruses to make broad-spectrum and effective drugs remained active in Russia. For example, using synthetic chemistry, bioinformatics-based analysis of S-protein, and molecular docking, new usnic acid derivatives with protease abilities were identified [176]. Another important avenue in the fight against COVID-19 was the development of rapid detection methods. In this regard, several research groups used ML methods for diagnosis using routine blood values [177] and rhythmogram [178] of patients. Therefore, through this example, it is understandable that the pharmaceutical and healthcare industries promoted research in the field of bioinformatics.

Agriculture's share in Russia's GDP has been increasing since 2000, and in 2020, it was 4.10%, up from 2014 (3.90%). Accounting for 8.8% of total exports (2020), agriculture is one of the key sectors considered to be the backbone of the Russian economy. The productivity index of the agriculture sector in 2020 was
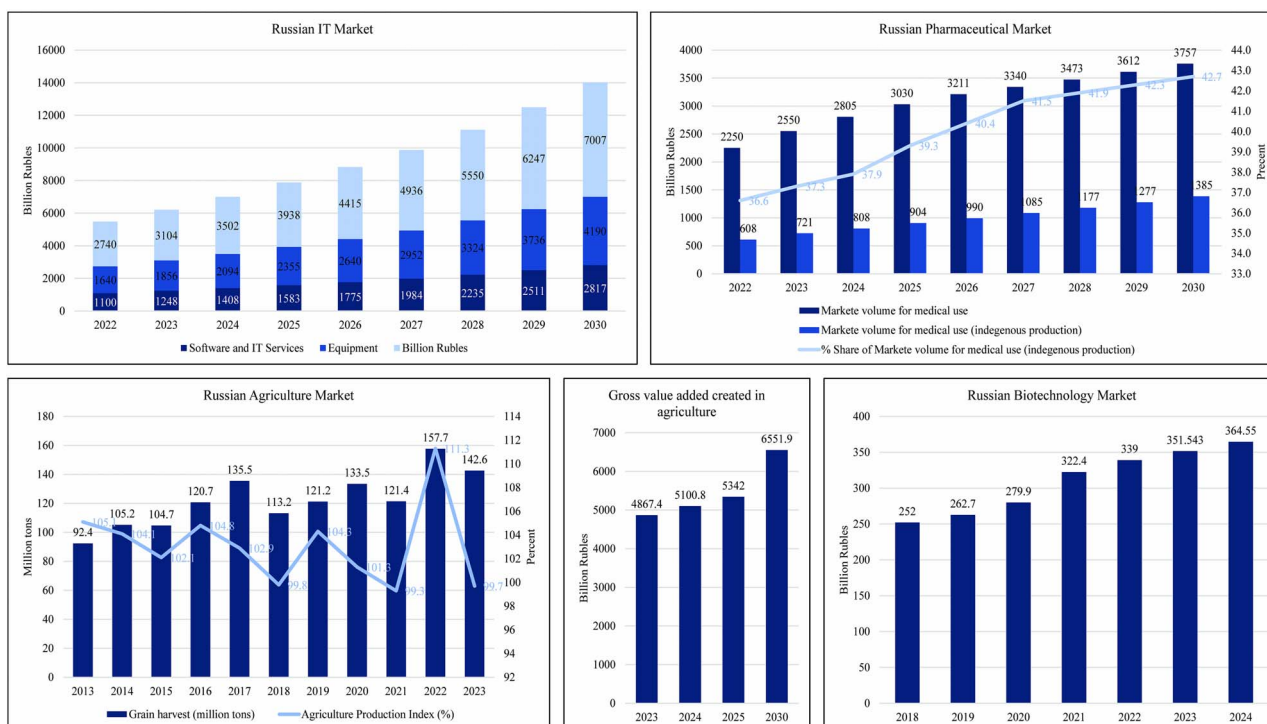
Figure 1. Value of four major Russian industries that contribute to the developments in bioinformatics; information technology [10], pharmaceuticals (source: government of the Russian Federation, distribution document number 1495-r dated June 7, 2023), agriculture (source: government of the Russian Federation, distribution document number 2567-r dated September 8, 2022), and biotechnology [11]. The 2023 and 2024 biotechnology market data were estimated based on the 2022 projections.

120.7%, which is higher than that of industry (108.6%) [22]. The Russian government has set the goal of increasing agricultural productivity by at least 25% by 2030 [23]. How to achieve this goal is a logical question since most of the registered crop varieties are developed either by classical selection or breeding methods (https://gossortrf.ru/registry/). The use of genome-wide association studies in plant breeding is just starting, e.g. loci associated with Zn, Mg, Mn, Ca, Cu, Fe, and K in the wheat grain of the Russian collection have been recently reported [24]. This deficiency is being recognized by the state as well as by plant breeders, and the focus is shifting to the use of genome sequencing. For example, genomic selection is underway for high protein content in wheat and soybean (https://www.rscf.ru/project/21-76-30003/). Some laboratories have initiated crop genome sequencing projects, e.g. WGS of 100 Russian-bred *Brassica oleracea* L. accessions to study associations with disease resistance by the Russian State Agrarian University (personal communication), WGS of 29 soybean accessions, and transcriptome sequencing of chickpea (https://www.skoltech.ru/research/en/gcf-2/). However, large-scale plant genome sequencing projects, such as the sequencing of 3366 chickpeas [25], 1000 soybeans [26], 3010 rice [27], etc., have not yet been realized. Lack of such large-scale projects (and data) might be a factor related to nonavailability and development of almost no plant omics analysis tools in Russia. Although some efforts are underway to the development of plant-specific genome databases, e.g. Genome Altai Plants (https://genome.asu.ru/).

The biotechnology sector itself is a driver for the development of the pharmaceutical, medical, food, and agricultural industries and has grown by 30% since 2015. The Program for the Development of Biotechnology in the Russian Federation up to 2020 (BIO 2020) had set goals that biotechnology production should be 3% of the gross domestic product (GDP) by 2030. Russia has also created the BioTech2030 platform for cooperation between education,

research, and industry in the field of biotechnology [28]. Recent reports by Innopraktika (https://innopraktika.ru/) and Abercade (http://abercade.ru/) suggest that the biotech sector has a higher production potential and its growth can significantly stimulate the development and production of bio-alternatives. The role of biotechnology in bioinformatics developments is visible from the fact that Russian scientists are mostly focused on metagenomics and prokaryotic genomics (see section 2.2). Russia is also actively involved in the research and use of clustered regularly interspaced short palindromic repeats (CRISPR)/Cas systems. Most of this is related to the development of semi-automated search systems, reduced-size Cas9 orthologues, creation of cellular models of human neurodegenerative diseases, etc. A detailed overview of CRISPR/Cas research can be found in the article by Shamkova et al. [29].

History of bioinformatics in Russia

Although bioinformatics sounds like a modern subject, it actually started with the development of the Edman degradation method for protein sequencing [30] and the determination of the amino acid sequence of insulin [31]. For a more detailed account of the history of bioinformatics from both biological and computer science perspectives, readers are referred to the brilliant article by Gauthier, Vincent [5]. Here, we present only the Russian perspective. Bioinformatics in Russia (and affiliated disciplines) can be traced back to the founding of the Academy of Sciences of the USSR (abbreviated as AH CCCP in Russian) by Pei Kib Breat in 1725. Although many important advances were made in the 1930 to 40s or earlier [32], the most recent history begins when George Gamow optimistically proposed a fairly precise genetic code in a letter to Watson and Crick [33], which stirred debate and initiated precision thinking in the field [34]. Pioneering work in the field of nucleic acids was done by Alexander Spirin and Andrei Belozersky,

who tested the correspondence between DNA and ribonucleic acid (RNA) [35]. They continued to work on mRNA and messenger ribonucleoprotein particles (informosomes) and discovered that the relocation of ribosomal subunits relative to each other is a key event in protein translation [34]. These were the times when somewhat parallel to the world [5], the applications of computers in different scientific fields also started in Russia [36]. This led to the involvement of several disciplines related to biology, such as mathematics. An important development in this aspect was the organization of the Institute of Mathematical Problems of Biology in 1972 (then known as the Scientific Research Computing Centre of the AH CCCP) (https://www.impb.ru). Major applications of computing began in the fields of genetics, microbiology, and molecular genetics and several conferences on bioinformatics were organized during 1970 to 80s [37].

Among the major developments in the field of bioinformatics, the growth of bioinformatics in Russia is marked by the establishment of key institutes and their subsidiaries. The first of these dedicated institutes is the Institute of Cytology and Genetics (ICG, founded in 1957) as a part of the Siberian Branch of the AH CCCP, which together with the Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of RAS (SB RAS) established SB RAS Genomics Core Facility in 2000 (http://www.niboch.nsc.ru/doku.php/corefacility; accessed on 22 April 2024). The second is the Institute for Information Transmission Problems of the RAS (1948; http://iitp.ru/en/about/history; accessed on 19 April 2024), which established a Research and Training Centre of Bioinformatics, Institute for Information Transmission Problems, RAS (RTCB) in 2003. Among other key institutes (Appendix 1), the Center for Collection Use "Bioinformatics" (CCUB, https://ckp.icgen.ru/sccbio/; accessed on 22 April 2024), created (in 2011) on the basis of the ICG of the SB RAS in cooperation with the Novosibirsk State University, is a major establishment. This center is also affiliated with the Centre for Collective Use "SSCC" which has a peak cluster performance of 115 TFLOPS. The SSCC and the CCUB have been actively involved in projects related to omics, molecular modeling, evolutionary bioinformatics, and molecular dynamics and have developed and used a number of packages, such as Gaussian and GROMACS [15].

Present day landscape of Russian bioinformatics

To clearly understand the direction of Russian bioinformatics, we searched the project catalog of the Russian Science Foundation (RSF, https://www.rscf.ru/en/; accessed on 23 April 2024), by using two keywords, i.e. "genomics" and "bioinformatics," alone or in combination against "All grant years" and "All research areas." This resulted in 198 research projects awarded to 73 institutions. By classifying the projects into 10 categories, a trend towards genetics (31.82%), OMICs (30.30%), medical informatics (27.78%), computational biology (16.16%), environmental informatics (7.07%), structural bioinformatics (4.55%), and drug discovery (3.54%) was observed (Fig. 2a). This highlights that biological modeling, machine-learning (ML), and biostatistics are the key areas for future focus, although biostatistics has existed in some form for the last 30 years. The largest number of projects were related to humans (OMICs, genetics, computational biology, and medical informatics), followed by plants (OMICs, genetics, drug discovery, and environmental informatics), animals (genetics), environmental informatics, and industry (Fig. 2b).

Another indicator of the developments in these fields is the number of publications in global databases such as PubMed. The number of publications categorized in the subjects of bioinformatics, genomics, and computational biology has increased over the last three decades (Fig. 3). However, a decrease in publications was observed from 2002 to 2012, suggesting that the progress was slower during this decade. Since 2014, however, an increase in publications has been observed, which aligns well with the establishment of the RSF. Although consistent with post-COVID-19 global trends, the decline in the number of publications after 2021 could also be due to the ongoing conflict between Russia and Ukraine [38]. Nevertheless, there are significant developments in the field of genomics, which might seem parallel to the global trend (Fig. 3).

Present day bioinformatics research in Russia

Modern bioinformatics began in Russia with the use of probabilistic approach(es) to solve the problems related to biomolecular linguistics [39], prediction of exonic structure of human genes [40], understanding of regulatory signals in bacterial genomes [41], DNA sequences [42], software for analysis of bacterial genomes [43], etc., which is parallel to the dawn of the genomics era in the last decade of the twentieth century [5]. Russia's bioinformatics is strongly allied with omics, computational biology, and the use of components of HPC for the development of data analysis tools and algorithms. On the omics front, there are a handful of institutes that are leading the developments in this field. The first of them is the RTCB, which focuses on comparative genomics, metabolic model reconstruction (Information Box-II), and the development of methods and algorithms for the identification of regulatory signals, alternative splicing, and 3D structures of proteins, RNA, and protein-DNA complexes.

Information Box-II

Transcription factors (TFs) recognize cognate TF-binding sites in the promoter regions of the regulons [179], which ultimately constitute the entire transcriptional regulatory network (TRN). Several approaches have been used for TRN reconstruction, including template-based, expression data-driven, and knowledge-driven comparative genomics [180]. Due to the advantages of knowledge-driven comparative genomics with the increasing availability of computational algorithms, genome context techniques, publicly available genome sequences, and National Microbial Pathogen Database Resource [127], it is now easier to reconstruct TRNs and identify new gene/TF members of regulons [181]. Researchers at RTCB have reconstructed TRNs of a large number of bacterial species belonging to phylogenetic groups of economic and environmental importance i.e. *Lactobacillacea* [182], *Bacillus subtilis* [183–185], *Shewanella* [186], *Firmicutes* [187], and *Thermotoga maritima* [188]. The work on *Clostridium* species (*Firmicutes*) for the reconstruction of xylose and xyloside utilization discovered a novel xylose isomerase [187]. This is applicable to the fermentation industry as these carbohydrates are abundant in cell-wall polymers of the hardwoods and crop residues. Other reconstructed TRNs or regulons include NiaR (YrxA) of the nicotinamide adenine dinucleotide metabolism [184], central carbon metabolism [189, 190], di-myo-inositol-phosphate [191], anaerobic respiration in gamma-proteobacteria [192], methionine/cysteine biosynthesis in *Corynebacterium glutamicum* [193], etc. Comparative analyses of up to 400 bacterial genomes for TRN reconstruction, e.g. metabolic pathways associated with glycerate-2-kinase [188], indicate that bacterial genomics is well established in Russia. RTCB researchers
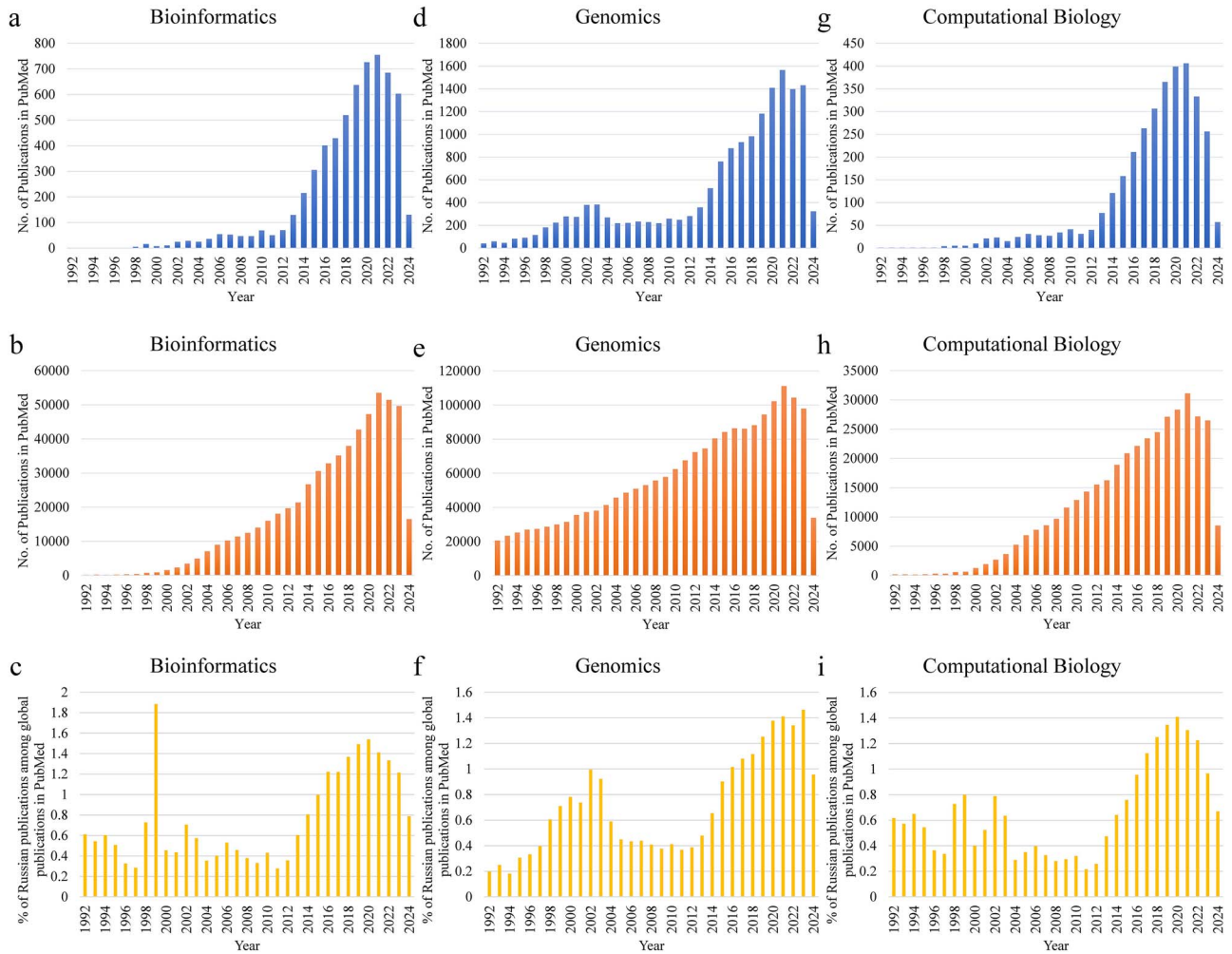
Figure 2. Classification and number of projects in the field of bioinformatics and affiliated disciplines funded by Russian Science Foundation (2014 to 2024). a) Number and % of projects according to subject classification. b) Number of projects dedicated to human, animal, plants, industrial, and environment related topics.
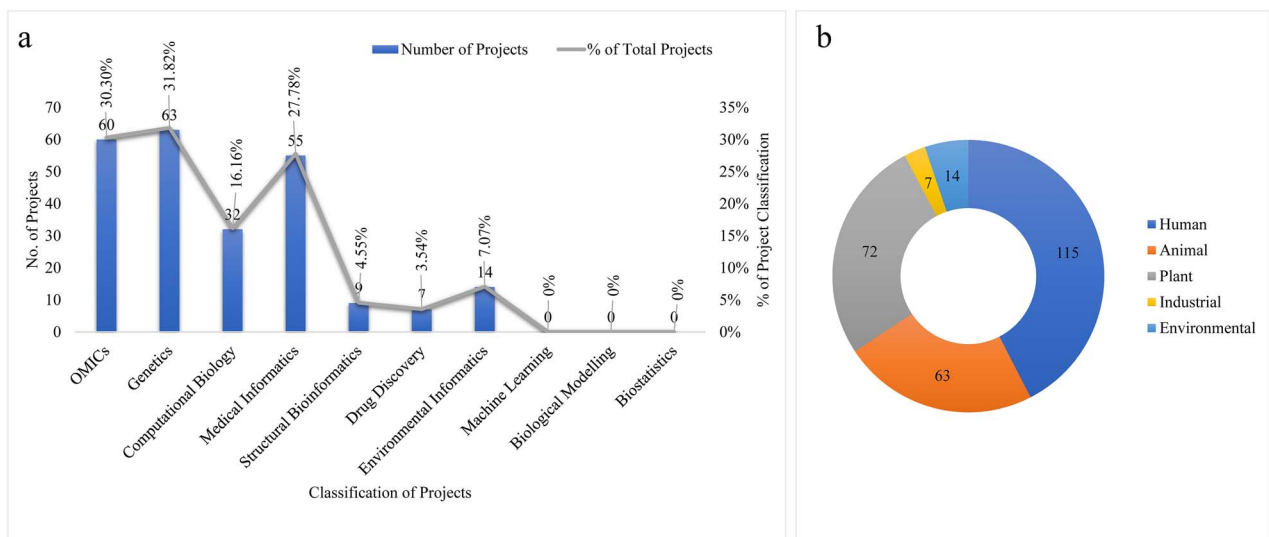

Figure 3. Number of bioinformatics, genomics, and computational biology publications in PubMed with Russian affiliation. Number of bioinformatics publications in a) Russia and b) world. c) % of Russian bioinformatics publications among bioinformatics' publications in the world. Number of genomics publications in d) Russia and e) world. f) % of Russian genomics publications among genomics' publications in the world. Number of computational biology publications in g) Russia and h) world. i) % of Russian computational biology publications among computational biology publications in the world.

have also developed a web resource based on these data for collection, visualization, and analysis of reconstructed regulons [143].

Among the other notable institutes, the SB RAS Genomics Core Facility has been a major contributor to developments in metagenomics and genomics of viruses, microbes, and eukaryotes; since 2005, the center has analyzed 555 740 samples. At the genomic level, the main contributions are the whole genome sequencing of bacteria (*Borrelia valaisiana* and *Borrelia afzelii* isolated from *Ixodes persulcatus* ticks [44]), *Coprobacter fastidiosus* NSB1T [45], chlamydia strain PP-87, *Bifidobacterium longum* subsp. longum Strain MC-42 [46], Thermaerobacter sp. strain PB12/ 4term [47], *Sinorhizobium meliloti* [48]), viruses (giant pseudomonas phage PaBG [49], Nucleopolyhedrovirus, Newcastle Disease Virus [50], H1N1 Swine Influenza Virus [51]), eukaryotes (*Lymantria dispar* [52], *Scleropages formosus*, *Lates calcarifer* [53], chloroplast sequence of *Pinus sibirica* Du Tour [54]). Most of these efforts report genome sequences rather than the development of tools/pipelines to solve related problems. However, these genome sequencing efforts also have practical applications. For example, the public is often reluctant to buy chicken raised with antibiotic additives. A nonantibiotic-pectin-based treatment has been developed that can modify the bacteriobiome of chickens [55]. Genome sequencing of bacteria isolated from ticks [44] proved to be an important step in understanding their control, as they are involved in the spread of the most common tick-borne infection (Lyme disease) in Eurasia. Genome sequencing of giant Pseudomonas phage PaBG indicated that it is phylogenetically distinct from other members of the *Myoviridae* family [49].

A recent addition to the field of bioinformatics is the Genomics Core Facility established at the Skolkovo Institute of Science and Technology, Moscow (2018). The facility is involved in using sequencing technologies to solve agricultural and health-related problems, such as comparative Illumina and Ion Torrent sequencing of herbal teas to establish and test DNA barcoding protocols [56]. An ongoing project to sequence 100 000 human genomes by 2025 marks the progress of bioinformatics in Russia. The center has completed the sequencing of 27 000 genomes (as of 2023) [57], which will ultimately enable researchers to develop diagnostics and therapies for genetic and socially important diseases. Some of the previous studies, involving 264 genomes and 615 individuals genotyped with Illumina BeadChips, reported six phylogeographic groups among the sampled ethnicities from selected Russian regions [58] and a genetic map of the genetic diversity of European Russia [59]. The institute has successfully launched projects to understand the causes of human pregnancy loss using exome analysis of abortuses, monogenic hereditary diseases, and how different concentrations of ouabain changes the gene expression profiles of human neurons (https://www.skoltech.ru/research/en/gcf-2/). A key test of the ability of Russian bioinformatics (an affiliated discipline) researchers to solve real-world problems on the genomics front is how they responded to the SARS-CoV-2 pandemic. In this respect, Russia's response was quite notable both on a national as well as global scale (Information Box 1). The institute is also active in enhancing knowledge of fundamental and applied aspects of prokaryotic and eukaryotic genomics e.g. role of neutral evolution in the violation of the triplet genetic code in *Euplotes ciliates* [60]. Moreover, the institute is also involved in developing algorithms e.g. solving problems related to removal of

biases in Hi-C data from multiple experiments [61] and prediction of chromatin folding [62].

From the perspective of agriculture and forestry, bioinformatics research is lagging behind, as most papers report on genome sequencing efforts but not on their utility in solving agricultural problems or on the development of data analysis tools/algorithms. However, some efforts are underway to use genome sequencing for association studies (as mentioned in section 2.1). Genomes (or assemblies) of *Vitis rotundifolia* Michx [63], Populus x sibirica [64], *Capsella bursa-pastoris* [65], *Prunus persica* cv. Sovetskiya [66], *Larix sibirica* Ledeb [67], *Fagopyrum esculentum* [68], fiber flax [69], etc., have been reported.

In addition to genome assembly, it is important to understand when and where these genes are expressed, which is achieved by transcriptome sequencing. Transcriptomics applications i.e. transcript assembly, variant detection, differential gene expression estimation, and further analysis provide useful insights. However, there are also computational challenges related to the nature and structure of transcripts and the amount of data generated [70]. To solve such problems, several research groups are actively developing pipelines such as trans2express-*de novo* transcriptome assembly pipeline [71], which produces a high-quality assembled genome, Cysmotif searcher pipeline for antimicrobial peptide identification in plant transcriptomes [72], and the ICAnnoLncRNA pipeline for LncRNA search and annotation for transcriptome seq data [73]. ICAnnoLncRNA uses LnchFinder and can successfully characterize lncRNAs from non-model species, which the earlier pipelines were unable to do [73]. In addition, several algorithms have been developed for transcriptome analysis of non-model organisms [74], RNA-seq data [75], and quality assessment of *de novo* transcriptome assemblies [76]. There are limited number of pipelines or algorithms that can combine both metagenomic and metatranscriptomic data. This problem has been solved by the development of the MetaGT pipeline [77]. Comparison of shotgun metagenomic data by reference-based, abstract-based, and neural network-based methods provide limited information and features are concealed. To address this, the MetaFast algorithm was developed, which uses adaptive segmentation of the metagenomic de Brujin graph [78]. During human embryonic development, it is crucial to track the origin of pediatric diseases. Research by Kameneva et al. revealed transitions that connect the intermediate mesoderm and progenitors of organ primordia, the hematopoietic system and endothelial subtypes in human embryos. By using single-cell transcriptomics, they successfully identified multiple sympathoblast lineages [79]. This technique is becoming increasingly popular as more and more research groups apply it to answer important questions. Gulaia et al. used single-nucleus transcriptomics of glioma tumor cells and highlighted that glioblastoma treatment must focus on the differentiation protocol of these cells [80]. Other research groups are studying immunity [81] and molecular mechanisms of human evolution [82]. The use of machine learning (ML) to compare interspecific gene expression profiles [83], clinically significant miRNAs from transcriptome data [84, 85], and the development of ML approaches for sequence prediction, such as TransPrise for eukaryotic promoters [86] and transcription start sites [87], mark the significant progress in combining modern ML approaches with omics [88]. These developments are mostly related to diseases associated with humans or metagenomic problems. Globally, transcriptomics has significantly improved the understanding of plant growth and development and stress responses [89]. However, in Russia, research on the plant transcriptomics is almost non-existent.

Some laboratories are working on transcriptomics of model plants (Arabidopsis) [90] and other (crop) plant transcriptomes e.g. barley [91], tomato [92], tea plant [93], and *Filipendula ulmaria* [94]. The use of long-read sequencing can help reveal complexity of plant transcriptomics [95]. Some Russian agriculturists are now using this technology to understand the landscape of plant development e.g, seed development in triticale [96]. Although it answers important questions e.g. how the plant retrotranscriptome originated [97], the development of tools and algorithms for plant-specific omics data analysis is almost non-existent, although a few exceptions are reported, e.g. NanoTRF pipeline for *de novo* satellite DNA identification from plant genomes [98].

Research at the protein level is well established in Russia, thanks to the work of Alexander Spirin and the organization of the Institute of Protein of the USSR Academy of Sciences in 1967 (now the Protein Institute of the RAS; https://protres.ru/; accessed on April 24, 2024). Starting with the development of the model of the ribosome function [99] and methods of peptide synthesis [100], protein research has progressed significantly. Here, we highlight the developments from the perspective of bioinformatics analyses, as explaining that the full scope of protein research in Russia is beyond the limits of this manuscript. At the level of collections and databases, our literature search produced limited results. A major focus of modern protein research is the prediction of protein structure and binding. Several protein-specific databases have been developed by individual research groups. A curated collection of binding models for human and mouse transcription factors (TFs), i.e. HOCOMOCO, has been developed [101]. Some researchers are also using molecular dynamics for protein studies. For example, a citrullinated protein associated with the development of rheumatoid arthritis was recently studied through molecular dynamics [102]. In order to design a desired protein/enzyme with the required activity, it may be necessary to change protein folding properties. Several patterns have been reported that allow the identification of amino acid residues that determine the folding rate and protein stability [103]. Meanwhile, protein structure prediction is already progressing around the world as databases with more accurate protein structure predictions are now available e.g. AlphaFold [104]. There are studies by Russian researchers reporting the correlation between folding and unfolding of proteins [105]. From the point of view of protein sequence, it is known that proteins evolve based on nucleotide substitutions in their corresponding genes. There is limited knowledge of how many different amino acid sequences a protein can acquire as a result of nucleotide substitutions. An algorithm has been proposed to study the evolutionary paths of sequences [106]. Several servers, algorithms, and search engines have been developed for other proteomic analyses, such as analysis of bulky MS/MS data in one night [107], regulated intramembrane proteolysis [108], antibody repertoire construction [109], and an extensible search engine for protein identification in shotgun proteomics [110]. Moreover, several groups have used proteomics-based ML approaches to diagnose diseases such as chronic kidney disease [111] and protein identification in single-shot human proteome analysis [112]. In the field of proteomics, Russia is a member of the Human Proteome Organization (HUPO) and a dedicated Center for Common Use "Human Proteome" (https://www.ibmc.msk.ru/megascience/proteocentre; accessed on 24 April 2024) has been operating for almost a decade. Almost all the research groups in Russia are active in applied proteomics. A Google Scholar search using the keywords "Russia" and "proteome" for the last five years (2019 to 2024) returned 17,200 results. Adding further keywords indicated that environment (17 500) and health (17 100) are the top

priority fields in proteomics followed by agriculture (14 700), and microbiology (9290). Regarding the use of proteomic data for the discovery of un-identified proteins, a meta-analysis of >1000 Russian proteome profiles helped to identify missing and uncertain proteins, and ~150 single proteotypic peptides [113]. Proteomics has also been successfully used to reveal ancient bone proteins of a Pleistocene mammoth [114], its diet [115], and its habitat [116]. At the level of protein families, the search and characterization of functional and regulatory ligand binding sites is carried out in several proteomics laboratories [117]. Applied proteomics in the field of agriculture is also gaining momentum in Russia. Research is underway on plant growth and development (proteome map of *Pisum sativum* embryos [118]), biotic stress (activation of proteases in early responses to biotic stress in wheat [119]), abiotic stress (extreme freezing tolerance in spruce [120]), and nutrition (proteomics of wheat pretreated with iron compounds [121]).

As more and more genomics and metagenomics projects are being initiated and completed, the provision of bioinformatics tools and software, as well as data resources and services is becoming an integral part of bioinformatics and allied disciplines. Analogs of the large omics databases of the National Centre for Biotechnology Information, the European Bioinformatics Institute, the National Genomics Data Centre, China, and the DNA Databank of Japan do not exist in Russia. However, there are several small-scale databases on genomics or related data. Russian scientists are aware of this lack, and a National Genomic Information Database (NGID) is being developed under the Federal Scientific and Technical Program for the Development of Genetic Technologies [122]. While the international databases like NCBI offer greater benefits in terms of data sharing, there are currently no forms of restrictions on uploading or downloading omics data. However, there are certain areas of convergence and divergence in data sharing under the ethical, national, and international approaches [123]. The existence of ethical and legal risks related to personal and genetic data protection and national legal acts governing such issues might be solved if a database is governed by national laws [124]. It would also be helpful to track the country's progress in the fields of bioinformatics and omics. Some examples of dedicated small-scale databases, omics analysis tools, and algorithms developed and curated in Russia are listed in Table 1.

Specialized degrees/diplomas in bioinformatics began in the mid-1990s in subjects related to mathematics, medical, and agricultural sciences [37, 149]. As of now, an increasing number of universities, research institutes, and centers in Russia have started teaching bioinformatics courses (Appendix 1). A major contribution to the development of bio-medical informatics discipline could be the establishment of Laboratory of Biomedical Informatics in 2005 based on St. Petersburg Institute of Informatics and Automation, RAS (https://biomedinfo.narod.ru/; accessed on 23 April 2024). However, the specialized bioinformatics degree programs were being taught at several universities by 2010. Bioinformaticians are in demand in Russia as more and more universities have or are starting to develop teaching modules and degrees in this subject. Moreover, classical biological research laboratories related to all life science disciplines have initiated the use of omics techniques, thus increasing the talent pool. This transition is either a result of the collaboration with international researchers or at an individual level. Additionally, the return of graduates who have completed their higher education and scientific training abroad is playing a significant role in the establishment of the subject. The growth of bioinformatics in Russia would lead to an increased demand for 'multi-talented' specialists with fundamental biological knowledge, modern genomics, and bioinformatic

Table 1. Examples of biological data resources, servers, and software packages developed and maintained in Russia

Name	Characteristic function	Link/URL/Reference
Omics resources and databases		
NGID	It will provide storage, integration and analysis of genetic data (under construction)	[122]
FBDGI	A data repository covers convicts genotypes	[125]
Biobank Russia	Open-source reference database for genome wide association studies	https://biobank.almazovcentre.ru/# [126]
RuSeq	A database for genetic variants	http://ruseq.ru/#/
BV-BRC	An integrated resource for bacterial and viral infectious disease research	https://www.bv-brc.org/ [127]
Proteomics of malignant cells	An interconnected information resource for proteomic and biomedical data	http://ef2.inbi.ras.ru
EROP-Moscow	An Information resource about structures and functions of natural oligopeptides	http://erop.inbi.ras.ru/
Victoria server of Bioengineering Center of RAS	A server hosting serval database related to genomic data (promoters, repeats, inversions, micro-satellites, etc.)	http://victoria.biengi.ac.ru/cgi-bin/indelper/index.cgi
SINErice	A database for SINE elements in rice genome	http://victoria.biengi.ac.ru/sinerice/
GTRD	A database that provides an integrated view of processed data of genes through user-friendly web interface.	http://gtrd.biouml.org/ [128]
Omics data analysis tools		
HiConfidence	Uncovering of the biological signal in Hi-C data affected by technical bias	[61]
MGSmodeller	Developed for generation, reconstruction, edition and calculation of molecular genetic networks models.	http://biomodelsgroup.ru/projects/ [129]
ICGenomics	Used to analyzes and visualizes DNA and protein sequences.	http://www-bionet.sscc.ru/icgenomics [130]
SPAdes	A toolkit for assembly and analysis of sequencing data.	https://github.com/ablab/spades/releases/tag/v3.15.5 [194]
Trans2express	A *de-novo* transcriptome assembly pipeline optimized for gene expression analysis.	http://83.149.250.217/trans2express/ [71]
rnaSPAdes	A *de-novo* transcriptome assembler used to compare its performance to existing tools.	[75]
QUAST and rnaQUAST	Tools for quality assessment of genome assemblies and *de-novo* transcriptome assemblies, respectively.	[76, 131]
BBMerge	Merging tool for paired-end shotgun sequence data.	[132]
IonHammer	Fast and accurate algorithm for IonTorrent read error correction.	[133]
UniPro UGENE	Combines bioinformatics tools in a single interface and allows to work with different biological data formats, visualization and to retrieve data.	https://ugene.net/ [134]
svist4get	Develops publication-quality figures by visualizing genomic signal tracks and transcriptome annotation at specific genomic locations.	https://bitbucket.org/artegorov/svist4get/
Protein structure discovery	A software package to computer proteomics tasks.	http://www-bionet.sscc.ru/psd/ [135]
Zebra	A web server for bioinformatics analysis of diverse protein families.	http://biokinet.belozersky.msu.ru/zebra [136]
Biosaur	An open-source Python software for liquid chromatography–mass spectrometry peptide feature detection with ion mobility support	https://github.com/abdrakhimov1/Biosaur [137]
IdentiPy	An extensible search engine for protein identification in shotgun proteomics.	[110]
Pyteomics	It is a cross-platform and open-source Python library providing a rich set of tools for MS-based proteomics.	[112]
FractionOpti-mizer	It is a method for optimal peptide fractionation in bottom-up proteomics.	http://hg.theorchromo.ru/FractionOptimizer [138]
DirectMS1Quant	Provide ultrafast quantitative proteomics with MS/MS-free mass spectrometry.	[139]
viQC	A visual and intuitive quality control for mass spectrometry-based proteome analysis	[140]

(Continued)

Table 1. Continued

Name	Characteristic function	Link/URL/Reference
Pipelines and algorithms		
MetaGT	A pipeline for *de novo* assembly of metatranscriptomes with the aid of metagenomic data.	https://github.com/ablab/metaGT [77]
GCDPipe	It uses known risk genes, cell type expression profiles, and drug gene-target data to classify risk gene.	[141]
ICAnnoLncRNA	A snake make pipeline for a long non-coding RNA search and annotation in transcriptome sequences.	[73]
ANANASTRA	A web server for the identification and annotation of regulatory single-nucleotide polymorphisms (SNPs) with allele-specific binding events.	https://ananastra.autosome.org/ [142]
RegPrecise 3.0	A database for capturing, visualization and analysis of transcription factor regulons.	https://regprecise.lbl.gov/ [143]
RegTransBase	It consists of two databases: one catalogs regulatory interaction and another housing curated transcription factor binding sites.	http://regtransbase.lbl.gov/ (the site couldn't be accessed as of April 24, 2024) [144]
DeepZ	It utilizes deep learning neural networks to integrate information from various data sources, such as DNA sequence properties, histone modification, etc., as well as it enables whole-genome annotation of Z-DNA.	[145]
HypercubeME	Data analysis from random mutagenesis experiments and helps in finding complex interaction between mutations in genes.	https://github.com/ivankovlab/HypercubeME.git [146]
Scavager	A versatile post search validation algorithm for shotgun proteomics based on gradient boosting,	https://bitbucket.org/markmipt/scavager [147]
PPLine	An automated pipeline for SNP, SAP, and splice variant detection in the context of proteogenomics.	[148]

skills. Some of the most recent examples are the opening of the Department of Bioinformatics and Mathematical Biology at Alferovsky University, initiation of Master of Science in Bioinformatics at several universities such as Labachevsky University (2020), National Research University Higher School of Economics (2021) and National Research Tomsk State University, Tomsk, Russia (in 2023). More recently, Moscow State University named after M.V. Lomonosov (Application of Machine Learning in Biology, 2024), Ulyanovsk State Pedagogical University (Bioinformatics, 2024), and Ulyanovsk Technical University (Bioinformatics, 2024) announced/started similar programs focused on bioinformatics and allied disciplines. The creation of new departments and laboratories in bioinformatics highlights two scenarios. First, the state of bioinformatics and related disciplines in Russia is primarily sufficient to initiate training in mass education. Second, the demand for bioinformaticians in Russia is growing. Given the demand, it is reasonable to ask about the salary status of bioinformaticians. Bioinformaticians in Russia can expect a monthly income of RUB 60 000 to 200 000 per month [150], which is less than in the United States of America (biological scientists and computer and information research scientists) 99 000 to 157 000 US$ per annum (https://www.bls.gov/oes/current/oes_nat.htm; accessed on 26 April 2024). However, in the absence of official figures, it is not logically possible to compare salaries with those of other countries.

These developments in bioinformatics are being driven by government foresight and policy-making. Some of the major recent events that illustrate the government's proactive approach to bioinformatics are discussed here. One of the key programs for the life sciences (including bioinformatics) is the Federal Scientific and Technical Program for the Development of Genetic Technologies (FSTP). Launched in 2019, this program was designed for eight years (until 2027) with the aim of solving the problems associated with accelerating the development of genetic technologies (gene editing, medicine, agriculture, industry, and prevention of biological emergencies). In 2022, the FSTP was extended to 2030, involving >100 research centers, laboratories, and universities. This, together with the development of a new tax code for the right-to-technology, the reduction of income tax, and revising appropriate development indicators, demonstrates a focused approach. An important part of any country's development is the implementation and application of scientific research in industry and ultimately, the improvement of people's lives. To this end, a mechanism was created (in 2022) for the rapid introduction of applied genetic technological results into industry (https://www.minobrnauki.gov.ru/action/fntp/; accessed on 29 April 2024 and http://kremlin.ru/acts/assignments/orders/67630; accessed on 29 April 2024). An important part of a country's research independence is a unified database for storing, sharing, and retrieving biological data. Recently, developments in this regard have been observed in China, where government and researchers are working together on such databases [151]. Though relatively on a smaller scale, India has also initiated Indian Biological Data Center (https://ibdc.dbtindia.gov.in/; accessed on 03 May 2024). This indicates that countries that are generating biological data are also establishing their indigenous databases [152]. The announcement of the completion of the NGID in 2025 and amendments in several related state regulations is a good step forward [122]. Another notable example is the creation of an information system on biological information under the jurisdiction of Federal Service for Supervision of Consumer Rights Protection and Human Welfare – Rospotrebnadzor in 2022 (post-COVID-19 measures). For this, regulations on the Federal State Information System for Sanitary and Epidemiological Information were amended so that genome of pathogens could be sequenced and data could be used for timely implementation

of anti-epidemic measures (https://rospotrebnadzor.ru/about/info/news_time/news_details.php?ELEMENT_ID=18907; accessed on 03 May 2024). In recent years, some collaborative efforts have been made to develop platforms that could bring together data scientists, bioinformaticians, and geneticists. For example, BioTools4You was announced in 2021 for joint work of experts in genetics and biotechnologies but the website address (https://biotools4you.com/; accessed on 03 May 2024) is not active. These announcements and continued amendments in the relevant legislation highlight the ongoing efforts at the government level to manage, store, and share biological datasets.

How far Russia has come in the global race of bioinformatics development is still an open question due to the unavailability of specific data on the value and volume of the industry. From 2023 to 2028, the global bioinformatics market will grow to USD 35.5 billion at a CAGR of 13.6%. This forecast is based on data from North America, Europe, Asia-Pacific, and the rest of the world and does not include Russia [153]. Most of the big companies involved in omics research are located outside Russia e.g. Thermo Fisher Scientific, Illumina, BGI, etc. Moreover, the sequencers, equipment, and reagents are mostly from international companies. In the first quarter of 2022, 80 to 85% of raw materials and reagents for laboratories, e.g. reagents for DNA analysis, hormones, blood tests, and sequencing kits, were imported from Southeast Asia and Europe [154]. This indicates the country's high dependence on foreign products, which has been disrupted by the ongoing Russia-Ukraine conflict. This has also proved to be an opportunity for the country to develop and improve indigenous markets and find alternatives. Nevertheless, at the time of writing of this article, there is limited to no data on the comparative analysis of the Russian bioinformatics market with other countries.

Future of bioinformatics in Russia and challenges

Bioinformatics in Russia is still gaining the status of an established discipline despite the country's rich history in protein and nucleic acid research. It is now catching up with the global standards in other affiliated disciplines, including ML, single-cell omics, multi-omics, system biology, and computational biology. At the policy-making level, the government has made and implemented important decisions to promote and strengthen this discipline in the country. The extension of the FSTP timeline to 2030 is indeed a good step, but an extended vision to 2050 should be established in light of the Sustainable Development Goals of the United Nations and Russia's Sustainable Development Goals [155]. It has been reported that by 2025, annual global genomic data acquisition and storage will be 1 Zetta-bytes and 2 to 40 exa-bytes, respectively [7]. What amount of this data is generated in Russia? Well, nobody knows the answer. To this end, the NGID initiative will ensure efficient management, analysis, visualization, and storage of biological data. Such efforts will enable governments and researchers to predict the direction of the research and decide on future policies. For example, at the time of writing this manuscript, limited information is available on how many genomes (plants, animals, humans, and microbes) have been sequenced in Russia. Moreover, the limited availability of a centralized national hub just like GitHub (https://github.com/) doesn't allow the authors to discreetly conclude on how many bioinformatics tools, websites, databases, algorithms, pipelines, etc., have been developed by Russian

programmers and bioinformaticians. China [151], Australia [156], and Germany [157] can be considered good examples, and national-level infrastructures are increasingly being developed. Progress in bioinformatics may highlight if the national biodiversity has been fully understood at genomic or transcriptomic level, creating opportunities to report local reference genomes. Since animal and plant breeders mostly use local gene pools, there is a need for genome sequencing of domestic breeds and varieties. Moreover, several plant, animal, and microbe species are regional; hence, determining their genomic landscape is only possible if one can comprehend what is yet to be sequenced and analyzed. Researchers from specific countries have been practicing this for some time now [158]. Another major factor that should be given due importance is the endangered species. Since countries and regions often have indigenous lists of endangered species, understanding their genomes would be a better help to local breeders and conservationists [159]. So far, it is somewhat understandable from the literature survey that relatively more focus has been given to prokaryotic genome sequencing rather than eukaryotes. Considering the role of agriculture in the Russian economy, the literature survey clearly indicates a lag. Future research in agriculture should focus on use of omics technologies for improving yield and associated traits in crop plants. At least for humans, local genome datasets can help finding solutions to the indigenous problems. Of some noticeable projects, "The Genome Russia project," was set to be a good example but it has not been realized yet [160] to a full extent. However, attempts have been made to analyze genome sequences for a fraction of the ethnic populations (264 individuals) of Russia [58]. The ongoing project to sequence 100 000 people by 2025 is highly valuable in this regard. Once this data is available, it will be necessary to focus on personalized medicine and the delivery of high-quality treatments to participants and those with similar genetics. Most importantly, it will be necessary to use the knowledge as part of the routine clinical care, i.e. to integrate genomics into Russian healthcare. How will this affect the ongoing integration of WGS into clinical practice? What additional knowledge would be generated, e.g. secondary pathogenic findings, predisposition to rare or common diseases, behavioral and psychological responses, etc., as gained from similar projects [161]? As a further step, RNome should be used to identify full range of expressed transcripts and associated variations in ethnic groups, groups with specific diseases or mutations, and to develop treatments and vaccines. The authors leave an open-ended question of whether such a project shall be launched and whether, as with HUPO, this effort should be undertaken in collaboration with the international community?

A key factor in the establishment of bioinformatics in the country is the development of new skills and the upskilling the existing pool of researchers. An increase in the amount of funding and the number of individuals or groups of researchers turning to bioinformatics is visible from the literature review. However, bioinformaticians will need to integrate wetlab experiments for functional validation of *in-silico* experiments, which are usually based on predictions and hypotheses. To this end, *ad hoc* workshops on the use of bioinformatics databases, data analysis, and visualization will need to be offered to bench biologists in ongoing formal degree programs and research centers projects. And for those bioinformaticians with mathematical and computational backgrounds, training and workshops on wet-lab techniques will be useful.

Key Points
- Four industries – information technology, pharmaceuticals, biotechnology, and agriculture - are the main drivers of bioinformatics research in Russia.
- A preliminary form of bioinformatics related to protein and nucleic acid research has existed in the country since the 1950s.
- Present research focuses on metagenomics, prokaryotic genomics, genetics, medical, environmental, and agricultural informatics, and computational biology.
- Russia lacks large-scale biological databases, but a NGID is being established.
- Future research should focus on molecular modeling, ML, biostatistics, eukaryotic genome sequencing, and database development.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Acknowledgements

Funding

References

1. Madhani PM. Indian bioinformatics: growth opportunities and challenges. *JMTP* 2011;**3**:7–17.
2. Pongor S, Landsman D. Bioinformatics and the developing world. *Biotechnol Dev Monit* 1999;**40**:10–3.
3. Venter JC, Adams MD, Myers EW. *et al.* The sequence of the human genome. *Science* 2001;**291**:1304–51. https://doi.org/10.1126/science.1058040.
4. Koch I, Fuellen G. A review of bioinformatics education in Germany. *Brief Bioinform* 2008;**9**:232–42. https://doi.org/10.1093/bib/bbn006.
5. Gauthier J, Vincent AT, Charette SJ. *et al.* A brief history of bioinformatics. *Brief Bioinform* 2019;**20**:1981–96. https://doi.org/10.1093/bib/bby063.
6. Bonetta L. Whole-genome sequencing breaks the cost barrier. *Cell* 2010;**141**:917–9. https://doi.org/10.1016/j.cell.2010.05.034.
7. Stephens ZD, Lee SY, Faghri F. *et al.* Big data: astronomical or genomical? *PLoS Biol* 2015;**13**:e1002195. https://doi.org/10.1371/journal.pbio.1002195.
8. Union IT. *Measuring Digital Develoment: The ICT Development Index 20232023*. ITU Telecommunication Development Bureau.
9. GlobalData. Russia ICT market size and forecast (by IT solution area, size band and vertical) to 2027. In: Center GI (ed.), *Global-Data Plc*. John Carpenter Street, London, United Kingdom, 2023.
10. Strategy Partners. *Russian Market of Infrastructure Software and its Development Prospects*. Business Center Space 1, Kompozitorskaya st. 17, Moscow, Russia. 2023 [cited 2024 09 August, 2024]; Available from: https://strategy.ru/research/research/47.
11. BussinessStat. *Analysis of the Biotechnology Market (BioTech) in Russia in 2018–2022, Forecast for 2023–2027*. Aronso business center, Elektrozavodskaya st., Moscow, Russia, 2022.
12. Fedorov AK, Gelfand MS. Towards practical applications in quantum computational biology. *Nat Comput Sci* 2021;**1**:114–9. https://doi.org/10.1038/s43588-021-00024-z.
13. Xu, Q. and Y. Zhang. *Research on the application of computer science in bioinformatics. Journal of Physics: Conference Series*. Glassfields, Avon Stree, Bristol, United Kingdom: IOP Publishing, 2021, 1915; 032042, https://doi.org/10.1088/1742-6596/1915/3/032042.
14. Glinsky BM. Siberian supercomputer center for collective use: stages of development, current state and prospects. *Problems of Computer Science* 2013;**3**:72–84.
15. Glinsky BM. *et al.* Supercomputer technologies in solving bioinformatics problems. *Software Systems: Theory and Applications* 2015;**6**:99–112.
16. Robinson T, Harkin J, Shukla P. Hardware acceleration of genomics data analysis: challenges and opportunities. *Bioinformatics* 2021;**37**:1785–95. https://doi.org/10.1093/bioinformatics/btab017.
17. Laboratory of Statistical Multiomics and Bioinformatics. *Development of High-Performance Computing Tools for Comprehensive Analysis of Multi-Omics Data and Development of Personalised Medicine*. [cited 2024 30.07.2024]; Available from: https://ibg.anrb.ru/naychnaya-deyatelnost/dejstvuyushhie-granty/laboratoriya-statisticheskoj-multiomiki-i-bioinformatiki/.
18. Noor A. Improving bioinformatics software quality through incorporation of software engineering practices. *PeerJ Comput Sci* 2022;**8**:e839. https://doi.org/10.7717/peerj-cs.839.
19. TADVISER. *Electronic Industry (Russian Market)*. Moscow, Russia: Central Federal District of the Russian Federation, 2024 [cited 2024 29.07.2024]; Available from: https://www.tadvisor.ru/index.php/Статья:Электронная_промышленность_(рынок_России).
20. The Observatory of Economic Complexity 5.0. *Pharmaceutical Products in Russia*. 2022 [cited 2024 17 July]; Available from: https://oec.world/en/profile/bilateral-product/pharmaceutical-products/reporter/rus.
21. Somda D, Wilson Kpordze S, Jerpkorir M. *et al.* In: Rudrapal M (ed). *The Role of Bioinformatics in Drug Discovery: A Comprehensive Overview*. IntechOpen Limited, Great Portland Street, London, United Kingdom: Drug Metabolism and Pharmacokinetics, 2023. https://doi.org/10.5772/intechopen.113712.
22. Emelyanova VA, Fedovsov VA. Agro-industrial complex of Russia in new geopolitical conditions. *Student Science Questions* 2023; **77**:5.
23. President of Russia. *Decree of the president of the Russian Federation of May 7, 2024 No. 309*. In: *On the National Development Goals of the Russian Federation for the Period up to 2030 and for the Perspective up to 2036*, 2024. https://kremlin.ru/events/president/news/73986.
24. Potapova NA, Timoshchuk AN, Tiys ES. *et al.* Multivariate genome-wide association study of concentrations of

seven elements in seeds reveals four new loci in Russian wheat lines. *Plants (Basel)* 2023;**12**:3019. https://doi.org/10.3390/plants12173019.

25. Varshney RK, Roorkiwal M, Sun S. *et al.* A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature* 2021;**599**:622–7. https://doi.org/10.1038/s41586-021-04066-1.

26. Bayer PE, Valliyodan B, Hu H. *et al.* Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. *Plant Genome* 2022;**15**:1–12. https://doi.org/10.1002/tpg2.20109.

27. Wang W, Mauleon R, Hu Z. *et al.* Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 2018;**557**:43–9. https://doi.org/10.1038/s41586-018-0063-9.

28. Abdrakhmanova GI, Belousov DR, Blinkin MYA. *et al.* Forecast of scientific and technological development of Russia: 2030. In: Gokhberg LM (ed). *Ministry of Education and Science of the Russian Federation and National Research University Higher School of Economics.* Myasnitskaya, 11, Moscow, Russia, p. 244, 2014.

29. Shmakova A, Shmakova OP, Karpukhina AA. *et al.* CRISPR/Cas: history and perspectives. *Russ J Dev Bio* 2022;**53**:272–82. https://doi.org/10.1134/S1062360422040075.

30. Edman P. A method for the determination of the amino acid sequence in peptides. *Arch Biochem* 1949;**22**:475–6.

31. Sanger F, Thompson E. The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *Biochem J* 1953;**53**:366–74. https://doi.org/10.1042/bj0530366.

32. Muzrukova EB. L.V.C.*Soviet Biology in the 30–40s: Crisis under the Conditions of a Totalitarian System*, Vol. **2**. Repressed Science: Universitetskaya nab., St. Petersburg, Russia: "Nauka", 1994;9:45–56.

33. Nanjundiah V. George Gamow and the genetic code. *Resonance* 2004;**9**:44–9. https://doi.org/10.1007/BF02903575.

34. Spirin AS. The ribosome as a conveying thermal ratchet machine. *J Biol Chem* 2009;**284**:21103–19. https://doi.org/10.1074/jbc.X109.001552.

35. Belozersky AN, Spirin AS. A correlation between the compositions of deoxyribonucleic and ribonucleic acids. *Nature* 1958;**182**:111–2. https://doi.org/10.1038/182111a0.

36. Fet YI, Pospelov DA. Parallel computing in Russia. In parallel computing technologies. In: Malyshkin VE (ed) *Third International Conference, PaCT-95 St. Petersburg.* Russia. September 12–25, 1995 Proceedings 3: Springer, 1995.

37. Fando R. The history of bioinformatic in Russia. In: *2020 International Conference Engineering Technologies and Computer Science (EnT).* Moscow, Russia: IEEE, 2020; 87–90. https://doi.org/10.1109/EnT48576.2020.00022.

38. Nazarovets M, da Silva JAT. *Scientific Publishing Sanctions in Response to the Russo-Ukrainian War.* Learned Publishing, 2022;**35**:658–70. https://doi.org/10.1002/leap.1487.

39. Gelfand MS, Kozhukhin CG, Pevzner PA. Extendable words in nucleotide sequences. *Bioinformatics* 1992;**8**:129–35. https://doi.org/10.1093/bioinformatics/8.2.129.

40. Gelfand M. Statistical analysis and prediction of the exonic structure of human genes. *J Mol Evol* 1992;**35**:239–52. https://doi.org/10.1007/BF00178600.

41. Vitreshchak A, Bansal AK, Titov II. *et al.* Computer analysis of regulatory signals in complete bacterial genomes. Translation initiation of ribosomal protein operons. *Biofizika* 1999;**44**:601–10.

42. Gel'fand M. Computer analysis of DNA sequences. *Mol Biol* 1998;**32**:103–20.

43. Mironov A, Vinokurova N, Gel'fand M. Software for analyzing bacterial genomes. *Mol Biol* 2000;**34**:253–62.

44. Kurilshikov AM, Fomenko NV, Stronin OV. *et al.* Complete genome sequencing of Borrelia valaisiana and Borrelia afzelii isolated from Ixodes persulcatus ticks in western Siberia. *Genome Announc* 2014;**2**:e01315–4. https://doi.org/10.1128/genomeA.01315-14.

45. Chaplin AV, Efimov BA, Khokhlova EV. *et al.* Draft genome sequence of Coprobacter fastidiosus NSB1T. *Genome Announc* 2014;**2**:00122–14. https://doi.org/10.1128/genomeA.00122-14.

46. Tupikin AE, Kalmykova AI, Kabilov MR. Draft genome sequence of the probiotic Bifidobacterium longum subsp. longum strain MC-42. *Genome Announc* 2016;**4**:01411–6.

47. Baturina OA, Pavlova ON, Novikova AS. *et al.* Draft genome sequence of *Thermaerobacter* sp. strain PB12/4term, a thermophilic facultative anaerobic bacterium from bottom sediments of Lake Baikal, Russia. *Microbiol Resour Announc* 2018;**7**:01178–18. https://doi.org/10.1128/MRA.01178-18.

48. Muntyan VS, Afonin AM, Vladimirova ME. *et al.* Complete genome sequence of Sinorhizobium meliloti S35m, a salt-tolerant isolate from alfalfa rhizosphere in soil native to the Caucasus region. *Microbiol Resour Announc* 2021;**10**:01417–20. https://doi.org/10.1128/MRA.01417-20.

49. Sykilinda NN, Bondar AA, Gorshkova AS. *et al.* Complete genome sequence of the novel giant pseudomonas phage PaBG. *Genome Announc* 2014;**2**:00929–13. https://doi.org/10.1128/genomeA.00929-13.

50. Glushchenko AV, Alikina TY, Yurchenko KS. *et al.* Nearly complete genome sequence of a Newcastle disease virus strain isolated from a wild garganey (spatula querquedula) in Russia. *Microbiol Resour Announc* 2019;**8**:01072–19. https://doi.org/10.1128/MRA.01072-19.

51. Sobolev I, Kurskaya O, Murashkina T. *et al.* Genome sequence of an unusual reassortant H1N1 swine influenza virus isolated from a pig in Russia, 2016. *Genome Announc* 2017;**5**:00747–17. https://doi.org/10.1128/genomeA.00747-17.

52. Kabilov MR, Martemyanov VV, Tupikin AE. *et al.* Complete genome sequence of a western Siberian Lymantria dispar multiple nucleopolyhedrovirus isolate. *Genome Announc* 2015;**3**:00335–15. https://doi.org/10.1128/genomeA.00335-15.

53. Vij S, Kuhl H, Kuznetsova IS. *et al.* Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. *PLoS Genet* 2016;**12**:e1005954. https://doi.org/10.1371/journal.pgen.1005954.

54. Baturina O, Tupikin AE, Goroshkevich SN. *et al.* The complete chloroplast genome sequences of Pinus sibirica Du tour. *Mitochondrial DNA Part B* 2019;**4**:286–7. https://doi.org/10.1080/23802359.2018.1542983.

55. Lysko SB, Baturina OA, Naumova NB. *et al.* No-antibiotic-pectin-based treatment differently modified cloaca bacteriobiome of male and female broiler chickens. *Agri* 2021;**12**:24. https://doi.org/10.3390/agriculture12010024.

56. Speranskaya AS, Khafizov K, Ayginin AA. *et al.* Comparative analysis of Illumina and ion torrent high-throughput sequencing platforms for identification of plant components in herbal teas. *Food Control* 2018;**93**:315–24. https://doi.org/10.1016/j.foodcont.2018.04.040.

57. GxPNews. *A Genome Database of 100 Thousand Russians Will be Created by the End of 2025.* Moscow: Russia, 2024. st. Barclay, 6 building 5, 121087. [Cited 2024 09 August 2024] Available from: https://gxpnews.net/en/2024/02/a-genome-database-of-100-thousand-russians-will-be-created-by-the-end-of-2025/.

58. Zhernakova DV, Brukhin V, Malov S. *et al.* Genome-wide sequence analyses of ethnic populations across Russia. *Genomics* 2020;**112**:442–58. https://doi.org/10.1016/j.ygeno.2019.03.007.

59. Khrunin AV, Khokhrin DV, Filippova IN. *et al.* A genome-wide analysis of populations from European Russia reveals a new pole of genetic diversity in northern Europe. *PLoS One* 2013;**8**:e58552. https://doi.org/10.1371/journal.pone.0058552.

60. Gaydukova SA, Moldovan MA, Vallesi A. *et al.* Nontriplet feature of genetic code in Euplotes ciliates is a result of neutral evolution. *Proc Natl Acad Sci* 2023;**120**:e2221683120. https://doi.org/10.1073/pnas.2221683120.

61. Kobets VA, Ulianov SV, Galitsyna AA. *et al.* HiConfidence: a novel approach uncovering the biological signal in hi-C data affected by technical biases. *Brief Bioinform* 2023;**24**:bbad044. https://doi.org/10.1093/bib/bbad044.

62. Rozenwald MB, Galitsyna AA, Sapunov GV. *et al.* A machine learning framework for the prediction of chromatin folding in *Drosophila* using epigenetic features. *PeerJ Comput Sci* 2020;**6**:e307. https://doi.org/10.7717/peerj-cs.307.

63. Agakhanov M, Grigoreva EA, Potokina EK. *et al.* Genome assembly of Vitis rotundifolia Michx. Using third-generation sequencing (Oxford Nanopore technologies). Труды по прикладной ботанике, генетике и селекции 2021;**182**:63–71. https://doi.org/10.30901/2227-8834-2021-2-63-71.

64. Melnikova NV, Pushkova EN, Dvorianinova EM. *et al.* Genome assembly and sex-determining region of male and female Populus× sibirica. *Front Plant Sci* 2021;**12**:625416. https://doi.org/10.3389/fpls.2021.625416.

65. Kasianov AS, Klepikova AV, Kulakovskiy IV. *et al.* High-quality genome assembly of Capsella bursa-pastoris reveals asymmetry of regulatory elements at early stages of polyploid genome evolution. *Plant J* 2017;**91**:278–91. https://doi.org/10.1111/tpj.13563.

66. Gladysheva-Azgari M, Petrova K, Tsygankova S. *et al.* A de novo genome assembly of cultivated Prunus persica cv. 'Sovetskiy'. *PLoS One* 2022;**17**:e0269284. https://doi.org/10.1371/journal.pone.0269284.

67. Kuzmin DA, Feranchuk SI, Sharov VV. *et al.* Stepwise large genome assembly approach: a case of Siberian larch (Larix sibirica Ledeb). *BMC Bioinformatics* 2019;**20**:35–46. https://doi.org/10.1186/s12859-018-2570-y.

68. Penin AA, Kasianov AS, Klepikova AV. *et al.* High-resolution transcriptome atlas and improved genome assembly of common buckwheat, Fagopyrum esculentum. *Front Plant Sci* 2021;**12**:612382. https://doi.org/10.3389/fpls.2021.612382.

69. Dmitriev AA, Pushkova EN, Novakovskiy RO. *et al.* Genome sequencing of fiber flax cultivar atlant using oxford nanopore and illumina platforms. *Front Genet* 2021;**11**:590282. https://doi.org/10.3389/fgene.2020.590282.

70. Mao S. *Computational Problems for RNA-Seq Data*. Analysis [Doctoral dissertation, University of Washington], Washington, United States of America, 2020. Available from: http://hdl.handle.net/1773/46333.

71. Kasianova AM. *et al.* Trans2express-de novo transcriptome assembly pipeline optimized for gene expression analysis. *Plant Methods* 2024;**20**:28. https://doi.org/10.1186/s13007-024-01255-7.

72. Shelenkov A, Slavokhotova A, Odintsova T. Cysmotif searcher pipeline for antimicrobial peptide identification in plant transcriptomes. *Biochemistry (Mosc)* 2018;**83**:1424–32. https://doi.org/10.1134/S0006297918110135.

73. Pronozin AY, Afonnikov DA. ICAnnoLncRNA: a snakemake pipeline for a long non-coding-RNA search and annotation in transcriptomic sequences. *Genes* 2023;**14**:1331. https://doi.org/10.3390/genes14071331.

74. Gubaev R, Gorshkov VY, Gapa LM. *et al.* Algorithm for physiological interpretation of transcriptome profiling data for non-model organisms. *Mol Biol* 2018;**52**:497–509. https://doi.org/10.1134/S0026893318040076.

75. Bushmanova E, Antipov D, Lapidus A. *et al.* rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* 2019;**8**:giz100. https://doi.org/10.1093/gigascience/giz100.

76. Bushmanova E, Antipov D, Lapidus A. *et al.* rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. *Bioinformatics* 2016;**32**:2210–2. https://doi.org/10.1093/bioinformatics/btw218.

77. Shafranskaya D, Kale V, Finn R. *et al.* MetaGT: a pipeline for de novo assembly of metatranscriptomes with the aid of metagenomic data. *Front Microbiol* 2022;**13**:981458. https://doi.org/10.3389/fmicb.2022.981458.

78. Ulyantsev VI, Kazakov SV, Dubinkina VB. *et al.* MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data. *Bioinformatics* 2016;**32**:2760–7. https://doi.org/10.1093/bioinformatics/btw312.

79. Kameneva P, Artemov AV, Kastriti ME. *et al.* Single-cell transcriptomics of human embryos identifies multiple sympathoblast lineages with potential implications for neuroblastoma origin. *Nat Genet* 2021;**53**:694–706. https://doi.org/10.1038/s41588-021-00818-x.

80. Gulaia V, Shmelev M, Romanishin A. *et al.* Single-nucleus transcriptomics of IDH1-and TP53-mutant glioma stem cells displays diversified commitment on invasive cancer progenitors. *Sci Rep* 2022;**12**:18975. https://doi.org/10.1038/s41598-022-23646-3.

81. Perik-Zavodskii R, Perik-Zavodskaia O, Shevchenko J. *et al.* Immune transcriptome study of human nucleated erythroid cells from different tissues by single-cell RNA-sequencing. *Cells* 2022;**11**:3537. https://doi.org/10.3390/cells11223537.

82. Khrameeva E, Kurochkin I, Han D. *et al.* Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Res* 2020;**30**:776–89. https://doi.org/10.1101/gr.256958.119.

83. Kasianov AS, Klepikova AV, Mayorov AV. *et al.* Interspecific comparison of gene expression profiles using machine learning. *PLoS Comput Biol* 2023;**19**:e1010743. https://doi.org/10.1371/journal.pcbi.1010743.

84. Mamoshina P, Volosnikova M, Ozerov IV. *et al.* Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front Genet* 2018;**9**:242. https://doi.org/10.3389/fgene.2018.00242.

85. Borisov N, Sergeeva A, Suntsova M. *et al.* Machine learning applicability for classification of PAD/VCD chemotherapy response using 53 multiple myeloma RNA sequencing profiles. *Front Oncol* 2021;**11**:652063. https://doi.org/10.3389/fonc.2021.652063.

86. Pachganov S, Murtazalieva K, Zarubin A. *et al.* TransPrise: a novel machine learning approach for eukaryotic promoter prediction. *PeerJ* 2019;**7**:e7990. https://doi.org/10.7717/peerj.7990.

87. Pachganov S, Murtazalieva K, Zarubin A. *et al.* Prediction of rice transcription start sites using TransPrise: a novel machine learning approach. In: Bandyopadhyay A, Thilmony R (eds) *Rice Genome Engineering and Gene Editing: Methods and Protocols*. Spring Streeet, New York, United States of America: Humana Press, pp. 261–74, 2021.

88. Tkachev V, Sorokin M, Borisov C. *et al.* Flexible data trimming improves performance of global machine learning methods in omics-based personalized oncology. *Int J Mol Sci* 2020;**21**:713. https://doi.org/10.3390/ijms21030713.

89. Vishwanath PP, Bidaramali V, Lata S. *et al.* Transcriptomics: illuminating the molecular landscape of vegetable

crops: a review. *J Plant Biochem Biotech* 2024;1–16. https://doi.org/10.1007/s13562-023-00871-z.

90. Doroshkov AV, Konstantinov DK, Afonnikov DA. *et al.* The evolution of gene regulatory networks controlling Arabidopsis thaliana L. trichome development. *BMC Plant Biol* 2019;**19**: 71–85. https://doi.org/10.1186/s12870-019-1640-2.

91. Shmakov NA. *et al.* Transcriptomic changes underlying partial albinism in barley nearly isogenic line. *PlantGen* 2019. https://doi.org/10.18699/PlantGen2019-169.

92. Nikolaychik EA, Gogoleva NE, Badalyan OA, Gogolev YV. Transcriptome analysis of *Solanum tuberosum* plant response to pectobacterial infection. *Cell Biology Plant Biotechnol, II International Sceintific-Practical Conference, Belarusian State University, Minsk, Republic of Belarus* 2018; P. 69. http://elib.bsu.by/handle/123456789/202330.

93. Samarina L, Matskiv AO, Koninskaya NG. *et al.* Comparative analysis of gene expression in tea plant (Camellia sinensis (L.) Kuntze) under low-temperature stress. *Vavilovskii Zhurnal Genet Selektsii* 2020;**24**:598–604. https://doi.org/10.18699/VJ20.653.

94. Istomina EA, Korostyleva TV, Kovtun AS. *et al.* Transcriptome-wide identification and expression analysis of genes encoding defense-related peptides of *Filipendula ulmaria* in response to *Bipolaris sorokiniana* infection. *J Fungi* 2024;**10**:4:258. https://doi.org/10.3390%2Fjof10040258.

95. Al-Dossary O, Furtado A, Kharabian Masouleh A. *et al.* Long read sequencing to reveal the full complexity of a plant transcriptome by targeting both standard and long workflows. *Plant Methods* 2023;**19**:112. https://doi.org/10.1186/s13007-023-01091-1.

96. Polkhovskaya E, Bolotina A, Merkulov P. *et al.* Long-read cDNA sequencing revealed novel expressed genes and dynamic transcriptome landscape of triticale (x Triticosecale Wittmack) seed at different developing stages. *Agronomy* 2023;**13**:292. https://doi.org/10.3390/agronomy13020292.

97. Kirov I, Merkulov P, Polkhovskaya E. *et al.* Epigenetic stress and long-read cDNA sequencing of sunflower (Helianthus annuus L.) revealed the origin of the plant Retrotranscriptome. *Plant (Basel)* 2022;**11**:3579. https://doi.org/10.3390/plants11243579.

98. Kirov I, Kolganova E, Dudnikov M. *et al.* A pipeline NanoTRF as a new tool for De novo satellite DNA identification in the raw nanopore sequencing reads of plant genomes. *Plant (Basel)* 2022;**11**:2103. https://doi.org/10.3390/plants11162103.

99. Spirin, A.S. A model of the functioning ribosome: Locking and unlocking of the ribosome subparticles. *Cold Spring Harb Symp Quant Biol* 1969;**34**:197–207. https://doi.org/10.1101/SQB.1969.034.01.026.

100. Mitin YV, Glinskaya O. Peptide synthesis using triphenyl phosphite and imidazole. *Tetrahedron Lett* 1969;**10**:5267–70. https://doi.org/10.1016/S0040-4039(01)88940-1.

101. Vorontsov IE, Eliseeva IA, Zinkevich A. *et al.* HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res* 2024;**52**:D154–63. https://doi.org/10.1093/nar/gkad1077.

102. Taldaev A, Rudnev V, Kulikova L. *et al.* Molecular dynamics study of Citrullinated proteins associated with the development of rheumatoid arthritis. *Proteomes* 2022;**10**:8. https://doi.org/10.3390/proteomes10010008.

103. Majorina M, Melnik TN, Glukhov AS. *et al.* Some useful ideas for multistate protein design: effect of amino acid substitutions on the multistate proteins stability and the rate of protein structure formation. *Front Mol Biosci* 2022;**9**:983009. https://doi.org/10.3389/fmolb.2022.983009.

104. Varadi M, Anyango S, Deshpande M. *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–44. https://doi.org/10.1093/nar/gkab1061.

105. Glyakina AV, Galzitskaya OV. How quickly do proteins fold and unfold, and what structural parameters correlate with these values? *Biomolecules* 2020;**10**:197. https://doi.org/10.3390/biom10020197.

106. Ivankov DN. Exact correspondence between walk in nucleotide and protein sequence spaces. *PloS One* 2017;**12**:e0182525. https://doi.org/10.1371/journal.pone.0182525.

107. Kiseleva O, Poverennaya E, Shargunov A. *et al.* Proteomic Cinderella: customized analysis of bulky MS/MS data in one night. *J Bioinform Comput Biol* 2018;**16**:1740011. https://doi.org/10.1142/S021972001740011X.

108. Ivankov DN, Bogatyreva NS, Hönigschmid P. *et al.* QARIP: a web server for quantitative proteomic analysis of regulated intramembrane proteolysis. *Nucleic Acids Res* 2013;**41**:W459–64. https://doi.org/10.1093/nar/gkt436.

109. Safonova Y, Bonissone S, Kurpilyansky E. *et al.* IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. *Bioinformatics* 2015;**31**:i53–61. https://doi.org/10.1093/bioinformatics/btv238.

110. Levitsky LI, Ivanov MV, Lobas AA. *et al.* IdentiPy: an extensible search engine for protein identification in shotgun proteomics. *J Proteome Res* 2018;**17**:2249–55. https://doi.org/10.1021/acs.jproteome.7b00640.

111. Glazyrin YE, Veprintsev DV, Ler IA. *et al.* Proteomics-based machine learning approach as an alternative to conventional biomarkers for differential diagnosis of chronic kidney diseases. *Int J Mol Sci* 2020;**21**:4802. https://doi.org/10.3390/ijms21134802.

112. Ivanov MV, Bubis JA, Gorshkov V. *et al.* Boosting MS1-only proteomics with machine learning allows 2000 protein identifications in single-shot human proteome analysis using 5 min HPLC gradient. *J Proteome Res* 2021;**20**:1864–73. https://doi.org/10.1021/acs.jproteome.0c00863.

113. Poverennaya E, Kiseleva O, Ilgisonis E. *et al.* Is it possible to find needles in a haystack? Meta-analysis of 1000+ MS/MS files provided by the Russian proteomic consortium for mining missing proteins. *Proteomes* 2020;**8**:12. https://doi.org/10.3390/proteomes8020012.

114. Cappellini E, Jensen LJ, Szklarczyk D. *et al.* Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins. *J Proteome Res* 2012;**11**:917–26. https://doi.org/10.1021/pr200721u.

115. Cucina A, Cunsolo V, di Francesco A. *et al.* Meta-proteomic analysis of the Shandrin mammoth by EVA technology and high-resolution mass spectrometry: what is its gut microbiota telling us? *Amino Acids* 2021;**53**:1507–21. https://doi.org/10.1007/s00726-021-03061-0.

116. Cucina A, di Francesco A, Saletti R. *et al.* Meta-proteomic analysis of two mammoth's trunks by EVA technology and high-resolution mass spectrometry for an indirect picture of their habitat and the characterization of the collagen type I, alpha-1 and alpha-2 sequence. *Amino Acids* 2022;**54**:935–54. https://doi.org/10.1007/s00726-022-03160-6.

117. Bezsudnova EY, Stekhanova TN, Suplatov DA. *et al.* Experimental and computational studies on the unusual substrate specificity of branched-chain amino acid aminotransferase from Thermoproteus uzoniensis. *Arch Biochem Biophys* 2016;**607**: 27–36. https://doi.org/10.1016/j.abb.2016.08.009.

118. Mamontova T, Lukasheva E, Mavropolo-Stolyarenko G. *et al.* Proteome map of pea (Pisum sativum L.) embryos containing different amounts of residual chlorophylls. *Int J Mol Sci* 2018;**19**:4066. https://doi.org/10.3390/ijms19124066.

119. Balakireva AV, Deviatkin AA, Zgoda VG. *et al.* Proteomics analysis reveals that caspase-like and metacaspase-like activities are dispensable for activation of proteases involved in early response to biotic stress in Triticum aestivum L. *Int J Mol Sci* 2018;**19**:3991. https://doi.org/10.3390/ijms19123991.

120. Kjellsen TD, Shiryaeva L, Schröder WP. *et al.* Proteomics of extreme freezing tolerance in Siberian spruce (Picea obovata). *J Proteomics* 2010;**73**:965–75. https://doi.org/10.1016/j.jprot.2009.12.010.

121. Kusainova TT, Emekeeva DD, Kazakova EM. *et al.* Ultra-fast mass spectrometry in plant biochemistry: response of winter wheat proteomics to pre-sowing treatment with iron compounds. *Biochemistry (Mosc)* 2023;**88**:1390–403. https://doi.org/10.1134/S0006297923090183.

122. Kolesnikov I, Nikolenko MV, Ermakov AV. *et al.* National Genomic Information Database. *Nanobiotechnology Reports* 2023;**18**:329–36. https://doi.org/10.1134/S2635167623700209.

123. Knoppers BM, Abdul-Rahman MNH, Bédard K. Genomic databases and international collaboration. *KLJ* 2007;**18**:291–311. https://doi.org/10.1080/09615768.2007.11427678.

124. Madej M, Karpiński P, Akopyan H. *et al.* Genetic data protection as an indispensable element of genomic medicine development. *Pol Arch Med Wewn* 2023;**133**:1–11. https://doi.org/10.20452/pamw.16425.

125. Попова Т, Сергеев А. Федеральная база данных геномной инϕормации в системе обеспечения баланса частных и публичных интересов в уголовном судопроизводстве. Юридическая наука и правоохранительная практика 2017;**39**:132–9.

126. Usoltsev D, Kolosov N, Rotar O. *et al.* Complex trait susceptibilities and diversity in a sample of 4,145 Russians. *Nat Commun* 2024;**15**:6212. https://doi.org/10.1038/s41467-024-50304-1.

127. McNeil LK, Reich C, Aziz RK. *et al.* The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. *Nucleic Acids Res* 2007;**35**:D347–53. https://doi.org/10.1093/nar/gkl947.

128. Kolmykov S, Yevshin I, Kulyashov M. *et al.* GTRD: an integrated view of transcription regulation. *Nucleic Acids Res* 2021;**49**:D104–11. https://doi.org/10.1093/nar/gkaa1057.

129. Kazantsev F, Akberdin I, Lashin S. *et al.* MGSmodeller–a computer system for reconstruction, calculation and analysis mathematical models of molecular genetic system. In: *Proceedings of the 6th International Conference on Bioinformatics of Genome Regulation and Structure*. Novosibirsk, Russia: Institute of Cytology and Genetics, 2008; p. 113.

130. Orlov YL, Bragin AO, Medvedeva I. *et al.* ICGenomics: a program complex for analysis of symbol sequences in genomics. *Vavilovskii Zhurnal Genet Selektsii* 2014;**16**:732–41.

131. Gurevich A, Saveliev V, Vyahhi N. *et al.* QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;**29**:1072–5. https://doi.org/10.1093/bioinformatics/btt086.

132. Bushnell B, Rood J, Singer E. BBMerge–accurate paired shotgun read merging via overlap. *PLoS One* 2017;**12**:e0185056. https://doi.org/10.1371/journal.pone.0185056.

133. Ershov V, Tarasov A, Lapidus A. *et al.* IonHammer: Homopolymer-space hamming clustering for IonTorrent read error correction. *J Comput Biol* 2019;**26**:124–7. https://doi.org/10.1089/cmb.2018.0152.

134. Okonechnikov K, Golosova O, Fursov M. *et al.* Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 2012;**28**:1166–7. https://doi.org/10.1093/bioinformatics/bts091.

135. Ivanisenko VA, Demenkov PS, Ivanisenko TV. *et al.* Protein structure discovery: a software package to computer proteomics tasks. *Bioorg Khim* 2011;**37**:17–29. https://doi.org/10.1134/S1068162011010080.

136. Suplatov D, Kirilin E, Takhaveev V. *et al.* Zebra: a web server for bioinformatic analysis of diverse protein families. *J Biomol Struct Dyn* 2014;**32**:1752–8. https://doi.org/10.1080/07391102.2013.834514.

137. Abdrakhimov DA, Bubis JA, Gorshkov V. *et al.* Biosaur: an open-source python software for liquid chromatography-mass spectrometry peptide feature detection with ion mobility support. *Rapid Commun Mass Spectrom* 2021;e9045. https://doi.org/10.1002/rcm.9045.

138. Solovyeva EM, Lobas AA, Kopylov AT. *et al.* FractionOptimizer: a method for optimal peptide fractionation in bottom-up proteomics. *Anal Bioanal Chem* 2018;**410**:3827–33. https://doi.org/10.1007/s00216-018-1054-2.

139. Ivanov MV, Bubis JA, Gorshkov V. *et al.* DirectMS1Quant: ultra-fast quantitative proteomics with MS/MS-free mass spectrometry. *Anal Chem* 2022;**94**:13068–75. https://doi.org/10.1021/acs.analchem.2c02255.

140. Solovyeva E, Lobas AA, Surin AK. *et al.* Viqc: visual and intuitive quality control for mass spectrometry-based proteome analysis. *J Anal Chem* 2019;**74**:1363–70. https://doi.org/10.1134/S1061934819140119.

141. Pinakhina D, Loboda A, Sergushichev A. *et al. Gene, cell type, and drug prioritization analysis suggest genetic basis for the utility of diuretics in treating Alzheimer disease.* Human genetics and genomics. *HGG Adv* 2023;**4**:100203. https://doi.org/10.1016/j.xhgg.2023.100203.

142. Boytsov A, Abramov S, Aiusheeva AZ. *et al.* ANANASTRA: annotation and enrichment analysis of allele-specific transcription factor binding at SNPs. *Nucleic Acids Res* 2022;**50**:W51–6. https://doi.org/10.1093/nar/gkac262.

143. Novichkov PS, Kazakov AE, Ravcheev DA. *et al.* RegPrecise 3.0–a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics* 2013;**14**:745–12. https://doi.org/10.1186/1471-2164-14-745.

144. Cipriano MJ, Novichkov PN, Kazakov AE. *et al.* RegTransBase–a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics* 2013;**14**:1–8. https://doi.org/10.1186/1471-2164-14-213.

145. Beknazarov N. and Poptsova M. DeepZ: a deep learning approach for Z-DNA prediction. In: Kim KK, Subramani VK (eds) *Z-DNA: Methods and Protocols*. Spring Streeet, New York, United States of America: Humana Press, 2023, p. 217–26. https://doi.org/10.1007/978-1-0716-3084-6_15.

146. Esteban LA Lonishin LR, Bobrovskiy D. *et al.* HypercubeME: two hundred million combinatorially complete datasets from a single experiment. *Bioinform* 2019;**36**:1960–62. https://doi.org/10.1093/bioinformatics/btz841.

147. Ivanov MV, Levitsky LI, Bubis JA. *et al.* Scavager: a versatile postsearch validation algorithm for shotgun proteomics based on gradient boosting. *Proteomics* 2019;**19**:e1800280. https://doi.org/10.1002/pmic.201800280.

148. Krasnov GS, Dmitriev AA, Kudryavtseva AV. *et al.* PPLine: an automated pipeline for SNP, SAP, and splice variant detection in the context of proteogenomics. *J Proteome Res* 2015;**14**:3729–37. https://doi.org/10.1021/acs.jproteome.5b00490.

149. Fando R, Klavdieva M. Bioinformatics: past and present. In: *2018 International Conference on Engineering Technologies and Computer Science (EnT)*. Moscow, Russia: IEEE, 2018;34–36. https://doi.org/10.1109/EnT.2018.00013.

150. Gridnev I. *What Bioinformaticians Do*. Limited Liability Company "Blastim" (ed), Habr Blockchain Publishing Ltd. Diagorou, Nicosia, Cyprus. 2023.

151. Wei L, Yu J. Bioinformatics in China: a personal perspective. *PLoS Comput Biol* 2008;**4**:e1000020. https://doi.org/10.1371/journal.pcbi.1000020.

152. Salter B, Zhou Y, Datta S. *et al.* Bioinformatics and the politics of innovation in the life sciences: science and the state in the United Kingdom, China, and India. *Science, Technology, & Human Values* 2016;**41**:793–826. https://doi.org/10.1177/0162243916631022.

153. BCC Research. *Bioinformatics: Technologies and global markets. BCC Publishing*. Boston: United States of America, 2023.

154. Expo Analytica. *The Events of February 2022 Certainly Had an Impact on the Reagent Segment*. Expo Analytica, Crocus Expo IEC, Moscow, Russia. 2022 [cited 2024 01.08.2024]; Available from: https://analitikaexpo.com/ru/about/news/2023/august/14/importozameshchenie-reagentov-v-rossii/.

155. Gukhman G. Vision of the future. Sustainable development goals of the Russian Federation until 2050. Энергия: экономика, техника, экология 2018;**10**:23–30. https://doi.org/10.31857/S023336190002440-0.

156. Ragan MA, Littlejohn T, Ross B. Genome-scale computational biology and bioinformatics in Australia. *PLoS Comput Biol* 2008;**4**:e1000068. https://doi.org/10.1371/journal.pcbi.1000068.

157. Tauch A, Al-Dilaimi A. Bioinformatics in Germany: toward a national-level infrastructure. *Brief Bioinform* 2019;**20**:370–4. https://doi.org/10.1093/bib/bbx040.

158. Thorburn DMJ, Sagonas K, Binzer-Panchal M. *et al.* Origin matters: using a local reference genome improves measures in population genomics. *Mol Ecol Resour* 2023;**23**:1706–23. https://doi.org/10.1111/1755-0998.13838.

159. Hogg CJ, Ottewell K, Latch P. *et al.* Threatened species initiative: empowering conservation action using genomic resources. *Proc Natl Acad Sci* 2022;**119**:e2115643118. https://doi.org/10.1073/pnas.2115643118.

160. Oleksyk TK, Brukhin V, O'Brien SJ. The genome Russia project: closing the largest remaining omission on the world genome map. *Gigascience* 2015;**4**:53. https://doi.org/10.1186/s13742-015-0095-0.

161. Nolan JJ, Forrest J, Ormondroyd E. Additional findings from the 100,000 genomes project: a qualitative study of recipient perspectives. *Genet Med* 2024;**26**:101103. https://doi.org/10.1016/j.gim.2024.101103.

162. Ma L, Li H, Lan J. *et al.* Comprehensive analyses of bioinformatics applications in the fight against COVID-19 pandemic. *Comput Biol Chem* 2021;**95**:107599. https://doi.org/10.1016/j.compbiolchem.2021.107599.

163. Burki T. First shared SARS-CoV-2 genome: GISAID vs virological. Org. *The Lancet Microbe* 2023;**4**:e395. https://doi.org/10.1016/S2666-5247(23)00133-7.

164. Kozlovskaya L, Piniaeva A, Ignatyev G. *et al.* Isolation and phylogenetic analysis of SARS-CoV-2 variants collected in Russia during the COVID-19 outbreak. *Int J Infect Dis* 2020;**99**:40–6. https://doi.org/10.1016/j.ijid.2020.07.024.

165. Tsimafeyeu I, Volkova M, Alekseeva G. *et al.* Safety and preliminary efficacy of the gam-COVID-vac vaccine and outcomes of SARS-CoV-2 infection in Russian patients with genitourinary malignancies. *J Hematol Oncol* 2021;**14**:1–7. https://doi.org/10.1186/s13045-021-01205-z.

166. Choo SZL, Teo SP. Gam-COVID-vac (sputnik V): a heterologous adenoviral vector-based COVID-19 vaccine. *Aging Pathobiol Ther* 2021;**3**:43–5. https://doi.org/10.31491/APT.2021.06.059.

167. Logunov DY, Dolzhikova IV, Shcheblyakov DV. *et al.* Safety and efficacy of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine: an interim analysis of a randomised controlled phase 3 trial in Russia. *Lancet Reg Health Eur* 2021;**397**:671–81. https://doi.org/10.1016/S0140-6736(21)00234-8.

168. Ryzhikov A. *et al.* A single blind, placebo-controlled randomized study of the safety, reactogenicity and immunogenicity of the "EpiVacCorona" vaccine for the prevention of COVID-19, in volunteers aged 18-60 years (phase I-II). Инфекция и иммунитет 2021;**11**:283–96.

169. Tukhvatulin AI, Dolzhikova IV, Shcheblyakov DV. *et al.* An open, non-randomised, phase 1/2 trial on the safety, tolerability, and immunogenicity of single-dose vaccine "sputnik light" for prevention of coronavirus infection in healthy adults. *The Lancet Regional Health–Europe* 2021;**11**:100241. https://doi.org/10.1016/j.lanepe.2021.100241.

170. Barchuk A, Bulina A, Cherkashin M. *et al.* Gam-COVID-vac, EpiVacCorona, and CoviVac effectiveness against lung injury during Delta and omicron variant surges in St. Petersburg, Russia: a test-negative case–control study. *Respir Res* 2022;**23**:276. https://doi.org/10.1186/s12931-022-02206-3.

171. Chistyakova G, Malgina GB, Ustyuzhanin AV. *et al.* Formation of anti-infectious and post-vaccination anti-SARS-CoV-2 humoral immunity in medical workers of the perinatal center. *Russ J Infect Immun* 2022;**12**:688–700. https://doi.org/10.15789/2220-7619-FOA-1856.

172. Team, V.G.C.V.T. *COVID19 Vaccine Tracker*. 2024 [cited 2024 30.07.2024]; Available from: https://covid19.trackvaccines.org/

173. Timofeeva AM, Sedykh SE, Sedykh TA. *et al.* Natural antibodies produced in vaccinated patients and COVID-19 convalescents recognize and hydrolyze oligopeptides corresponding to the S-protein of SARS-CoV-2. *Vaccines (Basel)* 2023;**11**:1494. https://doi.org/10.3390/vaccines11091494.

174. Timofeeva AM, Sedykh SE, Dmitrenok PS. *et al.* Identification of antibody-mediated hydrolysis sites of oligopeptides corresponding to the SARS-CoV-2 S-protein by MALDI-TOF mass spectrometry. *Int J Mol Sci* 2023;**24**:14342. https://doi.org/10.3390/ijms241814342.

175. Poolsup S, Zaripov E, Hüttmann N. *et al.* Discovery of DNA aptamers targeting SARS-CoV-2 nucleocapsid protein and protein-binding epitopes for label-free COVID-19 diagnostics. *Mol Ther Nucleic Acids* 2023;**31**:731–43. https://doi.org/10.1016/j.omtn.2023.02.010.

176. Yarovaya OI, Filimonov AS, Baev DS. *et al.* Usnic acid based thiazole-hydrazones as multi-targeting inhibitors of a wide spectrum of SARS-CoV-2 viruses. *New J Chem* 2023;**47**:19865–79. https://doi.org/10.1039/D3NJ03598K.

177. Velichko A, Huyut MT, Belyaev M. *et al.* Machine learning sensors for diagnosis of COVID-19 disease using routine blood values for internet of things application. *Sensors* 2022;**22**:7886. https://doi.org/10.3390/s22207886.

178. Stasenko SV, Kovalchuk AV, Eremin EV. *et al.* Using machine learning algorithms to determine the post-covid state of a person by their rhythmogram. *Sensors* 2023;**23**:5272. https://doi.org/10.3390/s23115272.

179. Moreno-Campuzano S, Janga SC, Pérez-Rueda E. Identification and analysis of DNA-binding transcription factors in Bacillus subtilis and other Firmicutes-a genomic approach. *BMC Genomics* 2006;**7**:1–10. https://doi.org/10.1186/1471-2164-7-147.

180. Rodionov DA. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem Rev* 2007;**107**: 3467–97. https://doi.org/10.1021/cr068309+.

181. Ravcheev DA, Best AA, Tintle N. *et al*. Inference of the transcriptional regulatory network in Staphylococcus aureus by integration of experimental and genomics-based evidence. *J Bacteriol* 2011;**193**:3228–40. https://doi.org/10.1128/JB.00350-11.

182. Ravcheev DA, Best AA, Sernova NV. *et al*. Genomic reconstruction of transcriptional regulatory networks in lactic acid bacteria. *BMC Genomics* 2013;**14**:1–14. https://doi.org/10.1186/1471-2164-14-94.

183. Leyn SA, Kazanov MD, Sernova NV. *et al*. Genomic reconstruction of the transcriptional regulatory network in Bacillus subtilis. *J Bacteriol* 2013;**195**:2463–73. https://doi.org/10.1128/JB.00140-13.

184. Rodionov DA, Li X, Rodionova IA. *et al*. Transcriptional regulation of NAD metabolism in bacteria: genomic reconstruction of NiaR (YrxA) regulon. *Nucleic Acids Res* 2008;**36**:2032–46. https://doi.org/10.1093/nar/gkn046.

185. Pinchuk GE, Rodionov DA, Yang C. *et al*. Genomic reconstruction of Shewanella oneidensis MR-1 metabolism reveals a previously uncharacterized machinery for lactate utilization. *Proc Natl Acad Sci* 2009;**106**:2874–9. https://doi.org/10.1073/pnas.0806798106.

186. Rodionov DA, Novichkov PS, Stavrovskaya ED. *et al*. Comparative genomic reconstruction of transcriptional networks controlling central metabolism in the Shewanella genus. *BMC Genomics* 2011;**12**:S3–17. https://doi.org/10.1186/1471-2164-12-S1-S3.

187. Gu Y, Ding Y, Ren C. *et al*. Reconstruction of xylose utilization pathway and regulons in Firmicutes. *BMC Genomics* 2010;**11**: 1–14. https://doi.org/10.1186/1471-2164-11-255.

188. Yang C, Rodionov DA, Rodionova IA. *et al*. Glycerate 2-kinase of Thermotoga maritima and genomic reconstruction of related metabolic pathways. *J Bacteriol* 2008;**190**:1773–82. https://doi.org/10.1128/JB.01469-07.

189. Leyn SA, Li X, Zheng Q. *et al*. Control of proteobacterial central carbon metabolism by the HexR transcriptional regulator: a case study in Shewanella oneidensis. *J Biol Chem* 2011;**286**: 35782–94. https://doi.org/10.1074/jbc.M111.267963.

190. Goelzer A, Bekkal Brikci F, Martin-Verstraete I. *et al*. Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of Bacillus subtilis. *BMC Syst Biol* 2008;**2**:1–18. https://doi.org/10.1186/1752-0509-2-20.

191. Rodionov DA, Kurnasov OV, Stec B. *et al*. Genomic identification and in vitro reconstitution of a complete biosynthetic pathway for the osmolyte di-myo-inositol-phosphate. *Proc Natl Acad Sci* 2007;**104**:4279–84. https://doi.org/10.1073/pnas.0609279104.

192. Ravcheev DA, Gerasimova AV, Mironov AA. *et al*. Comparative genomic analysis of regulation of anaerobic respiration in ten genomes from three families of gamma-proteobacteria (Enterobacteriaceae, Pasteurellaceae, Vibrionaceae). *BMC Genomics* 2007;**8**:1–17. https://doi.org/10.1186/1471-2164-8-54.

193. Kovaleva GY, Gelfand M. Regulation of methionine/cysteine biosynthesis in Corynebacterium glutamicum and related organisms. *Mol Biol* 2007;**41**:126–36. https://doi.org/10.1134/S0026893307010177.

194. Bankevich A, Nurk S, Antipov D. *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77. https://doi.org/10.1089/cmb.2012.0021.