Original Study Article
https://doi.org/10.36233/0372-9311-704

Criteria for assessment of the quality of *Pseudomonas aeruginosa* genome sequences

Alexey A. Kovalevich✉, Alexey S. Vodopianov, Ruslan V. Pisanov

Rostov-on-Don Antiplague Scientific Researsh Institute, Rostov-on-Don, Russia

Abstract

Introduction. With the development of sequencing technologies, the volume of genomic data is increasing, which necessitates the development of metrics for assessing the quality of genome assembly. Despite the unified nature of modern instruments (Plantagora, SQUAT, QUAST, BUSCO, CheckM2, etc.), they do not take into account the specific genome organization of particular species. The issue of import substitution of bioinformatics tools is particularly acute given limited access to foreign technologies. Furthermore, there are no specialized methods for assessing the quality of *Pseudomonas aeruginosa* genome assemblies, which is limited to general metrics (N50, number of contigs).

The aim of the study is to develop an algorithm and criteria based on a comprehensive approach for the specific assessment of the quality of whole-genome sequencing of *P. aeruginosa*.

Materials and methods. The study was conducted on 108 strains of *P. aeruginosa*. The proprietary software is developed in Java and Python languages.

Results. An algorithm for assessing the quality of *P. aeruginosa* whole-genome data has been developed based on the analysis of key housekeeping genes (*fur, algU, dinB*, etc.), genome size, GC content, and the N50 value. Genomes lacking key genes or with structural errors are classified as poor or medium, with the latter not recommended for phylogenetic analysis. The algorithm offers simple and clear parameters for assessing the quality of whole-genome data.

Conclusion. Based on the analysis of essential genes, genome size, GC content, and the N50 index, we have developed a classification of the quality of *P. aeruginosa* genome assemblies (good, medium, low). An algorithm and the Genomes Validator program have been created for rapid assessment.

Keywords: *Pseudomonas aeruginosa, whole-genome sequencing, housekeeping genes, quality assessment*

Funding source. The study was conducted as part of the Rospotrebnadzor industry research program (2021–2025).

Conflict of interest. The authors declare no apparent or potential conflicts of interest related to the publication of this article.

For citation: Kovalevich A.A., Vodopianov A.S., Pisanov R.V. Criteria for assessment of the quality of *Pseudomonas aeruginosa* genome sequences. *Journal of microbiology, epidemiology and immunobiology.* 2025;102(5):583–591.
DOI: https://doi.org/10.36233/0372-9311-704
EDN: https://www.elibrary.ru/IESFVC

Оригинальное исследование
https://doi.org/10.36233/0372-9311-704

Критерии оценки качества геномов *Pseudomonas aeruginosa*

Ковалевич А.А.✉, Водопьянов А.С., Писанов Р.В.

Ростовский-на-Дону ордена Трудового Красного Знамени научно-исследовательский противочумный институт Роспотребнадзора, Ростов-на-Дону, Россия

Аннотация

Введение. С развитием технологий секвенирования растёт объём геномных данных, что требует разработки показателей для оценки качества сборок геномов. Современные инструменты (Plantagora, SQUAT, QUAST, BUSCO, CheckM2 и др.) являются унифицированными, но при этом не учитывают особенностей организации генома конкретных видов. Особенно остро стоит вопрос импортозамещения биоинформационных инструментов в условиях ограниченного доступа к зарубежным технологиям. Кроме того, отсутствуют специализированные методы оценки качества сборок генома *Pseudomonas aeruginosa*, что ограничивается общими метриками (N50, количество контигов).

Цель работы — разработка алгоритма и критериев на основе комплексного подхода для специфической оценки качества полногеномного секвенирования представителей вида *P. aeruginosa*.

Материалы и методы. Исследование проводили на 108 штаммах *P. aeruginosa*. Авторское программное обеспечение разработано на языках Java и Python.
Результаты. Разработан алгоритм оценки качества полногеномных данных *P. aeruginosa* на основе анализа ключевых генов жизнеобеспечения (*fur, algU, dinB* и др.), размера генома, GC-состава и показателя N50. Геномы с отсутствием ключевых генов или структурными ошибками классифицируются как плохие или средние, последние не рекомендуются для филогенетического анализа. Алгоритм предлагает простые и понятные параметры оценки качества полногеномных данных.
Заключение. На основе анализа генов жизнеобеспечения, размера генома, GC-состава и показателя N50 нами разработана классификация качества сборок геномов *P. aeruginosa* (хорошее, среднее, низкое). Созданы алгоритм и программа «Genomes Validator» для оперативной оценки.

Ключевые слова: *Pseudomonas aeruginosa, полногеномное секвенирование, гены жизнеобеспечения, оценка качества*

Источник финансирования. Исследование проведено в рамках отраслевой научно-исследовательской программы Роспотребнадзора (2021–2025).

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Для цитирования: Ковалевич А.А., Водопьянов А.С., Писанов Р.В. Критерии оценки качества геномов *Pseudomonas aeruginosa*. *Журнал микробиологии, эпидемиологии и иммунобиологии.* 2025;102(5):583–591.
DOI: https://doi.org/10.36233/0372-9311-704
EDN: https://www.elibrary.ru/IESFVC

Introduction

With the development of high-throughput sequencing technologies and the decrease in their cost, the volume of genomic data produced is growing exponentially. Projects using large datasets of whole-genome sequencing (WGS) have many advantages: statistical power is increased, and it becomes possible to test various hypotheses about the micro- and macroevolution of genomes.

The continuous improvement of sequencing technologies and bioinformatics analysis has increased the significance of WGS in biology, medicine, pharmaceuticals, and agriculture, stimulating comparative genomic research. However, the growth in the number of sequencing projects and laboratories has led to an increase in the number of genome assemblies that are not always suitable for analysis. This highlighted the need to evaluate the quality of whole genome assembly data for researchers, who use it. This in turn created a necessity to develop standard metrics for comparing the quality of genome assemblies and annotations, as well as for evaluating the effectiveness of different methods used to obtain them.

Recent studies on genome assembly quality assessment have focused either on pre-assembly quality control or on the assembly evaluation in terms of contiguity and correctness. However, the assessment of correctness depends on the reference and is not applicable to *de novo* assembly projects. Therefore, it is worth studying methods that allow for quality assessment reports to be obtained both after and before assembly, to check the quality/correctness of *de novo* assembly and input data [1].

For genome assemblies, metrics such as the number of contigs, the number of scaffolds, and N50 (the maximum contig's length at which the total length of all contigs no shorter than this value accounts for at least 50% of the total length of all contigs in the assembly) provide only a brief overview of genome quality, not always reflecting its analytical suitability.

In turn, there are currently a sufficient number of resources and methods for the post-analysis stage of work, as well as for assessing genome quality: Picard [2], SQUAT [1], Plantagora [3], QUAST [4], CheckM1 [5], CheckM2 [6], GenomeQC [7], BUSCO [8]. However, they are unified and represent algorithms with different orientations, sometimes suitable for analyzing only eukaryotic organisms, while not taking into account the specific genome organization of a particular species. One of the most versatile and widely used instruments that utilizes genes to assess WGS data is BUSCO. Unlike the solutions mentioned above, BUSCO focuses on genome analysis using evolutionarily conserved orthologous genes, which are considered universal for certain taxonomic groups (bacteria, fungi, plants or animals). However, BUSCO does not provide an answer about the quality of the analyzed genome, only indicating the percentage of found/not found orthologous genes, and the final conclusion must be drawn by the specialist themselves. However, orthologous genes can be lost without affecting bacterial viability, unlike housekeeping genes, which can lead to an underestimation of genome quality.

Currently, WGS of infectious disease pathogens is widely used to study them, determine their origin, and track their spread. To assess the quality of such a large amount of data, domestic software tools are necessary. The latter is particularly important given that import substitution is becoming one of the strategic objectives in conditions where access to foreign technologies and foreign databases is difficult [9].

There are currently no evaluation criteria for WGS data for *Pseudomonas aeruginosa*. There are software

services that perform assessments based on general (non-specific) criteria (N50, number of contigs, etc.) and do not take into account the characteristics of a specific microorganism.

The aim of the study is to create an algorithm and criteria for assessing the quality of WGS data from *P. aeruginosa* representatives, as well as to develop a domestic software capable of evaluating the quality of WGS data.

Materials and methods

The study used 108 genomes of *P. aeruginosa* strains: 24 strains were obtained from the Collection of Pathogenic Microorganisms of the Rostov-on-Don Antiplague Institute of Rospotrebnadzor (isolated in Rostov-on-Don, Khabarovsk, and Mariupol in 2022–2024), and 84 strains were obtained from the international NCBI database. WGS was conducted as part of the implementation of the federal project for the socio-economic development of the Russian Federation until 2030, "Sanitary Shield of the Country — Health Security (Prevention, Detection, Response)". Sequencing was performed on the MiSeq platform (Illumina) using the MiSeq Reagent Kit v2 (500-cycles) (Illumina). This method allows for reads 2 × 251 nucleotides long, with genome coverage ranging from 8 to 20.

The assessment of the primary sequencing data was performed using the FastQC program. The collected WGS data was analyzed using the QUAST program [4, 10]. The Trimmomatic [11] and Lighter [12] algorithms were used for trimming and for reads correction. Genome assembly from reads was performed using the Spades program [13]. All genomes have passed an initial assessment using the Kraken 2 program, which allows for the identification of DNA fragment belonging to various prokaryotic species [14]. The WGS data of strain PAO1 from the international NCBI database [15] were used as a reference genome.

The proprietary software was developed in the Java and Python programming languages. The algorithm for searching for gene sequences in the assembly was performed with the use of Smith-Waterman local alignment with a minimum similarity threshold of 80%.

The confidence interval was calculated, differences were considered significant at $p < 0.05$.

Results

It is known that the genome of the causative agent of *Pseudomonas aeruginosa* infection contains a number of genes that are critical for its viability. These genes are called housekeeping genes. It is evident that if any of these genes are missing from the WGS data, it is a sequencing and/or genome assembly error. This very feature was the basis of our proposed algorithm – all essential genes should be detected in a good-quality sequence. Of course, the selection of genes to be used for quality control is of great importance in this process.

One of the criteria we devised for the algorithm to assess genome quality is the selection of genes based on the following criteria:

- nucleotide sequences must be within 1000 bp;
- the gene must be a single-copy;
- the gene must be directly involved in the microorganism's physiological activity, performing essential functions for its life processes;
- the gene must present in all strains of *P. aeruginosa*.

The *oprI* gene was chosen for rapid species identification of *P. aeruginosa*. The main task is to assess the quality of sequencing data not only based on the identification of essential genes but also on translating their sequences. Taking in account that these genes are critical for the existence of a microbial cell, their absence from the genome or critical translation errors (stop codons) are considered as sequencing errors.

Housekeeping genes used for validating the selected whole-genome sequences of *P. aeruginosa*: *fur, algU, dinB, dnaQ, holA, holB, PA0472, fpvI, tonB1, cntL, sigX, capB, cspD, groES, rpoH*. The selected genes are essential for functioning and survival in the environment and in a macroorganism. The following parameters were chosen as criteria for evaluating the whole-genome sequences: the GC content of the *P. aeruginosa* genomes, the size of the *P. aeruginosa* whole-genome sequence, and the N50 scaffold value.

After conducting the research and selecting the quality assessment criteria for genomes, the Genomes Validator software was developed, which, for convenience, operates in "batch mode," analyzes an unlimited number of genomes, and presents the results in tabular form. For each genome, the original file name, species, quality (poor, average, good), length, N50 value, GC content, as well as the reason for the invalidity of the genome are indicated (**Fig. 1**).

The developed program Genomes Validator is a cross-platform, which has a graphical interface, does not require installation, allows to analyze multiple genomes at the same time, and is available for downloading at https://github.com/alexeyvod/GenomesValidator. It has an intuitive interface and is user-friendly for those without programming skills.

The program was validated on a sample of 108 whole-genome sequences of *P. aeruginosa* strains. Following validation, genomes of good (63%), medium (29%), and poor (8%) quality were identified. Further analysis identified 37% of the genomes analyzed (of medium and poor quality), which can help avoid errors in subsequent calculations using bioinformatics methods. The average N50 value among the sample was 1,250,527.

The parameters N50, genome length, and GC content were identical to the values obtained from the programs used for comparison: CheckM2 and QUAST. However, these programs do not provide genome quality assessment metrics.

DOI: https://doi.org/10.36233/0372-9311-704

When assessing the quality of bacterial genomes using CheckM2 software, we found out that genomes with a Completeness value of 100 showed significant variability in the Contamination index. At the same time, the use of the Genomes Validator made it possible to estimate additionally the size of the whole-genome sequence, which may be more informative for practical analysis of WGS data. (**Fig. 2**). This parameter allows for a preliminary assessment of the presence of extrachromosomal elements in the genome of the strain under study. It should be noted that contamination with foreign DNA usually affects the overall GC content and causes a significant change in genome size, whereas the presence of plasmids or other mobile genetic elements does not lead to significant changes in this parameter.

Based on the statistical analysis of the Completeness and Contamination parameters (**Table**), it was found that Contamination values in the range of 2 to 8 may indicate a possible low reliability of the obtained WGS data. However, such results could also be due to specific characteristics of the clinical isolate's genome. Thus, the genome of strain Ps-agn-2889, analyzed in the CheckM2 program, has a Completeness score of 100 with a Contamination score of –35.65, but the reason for the contamination is not clear from the data obtained. Analysis in the program Genomes Validator revealed a 1.5-fold increase in genome size and GC content, indicating clear contamination with foreign bacterial DNA. The genome of clinical strain 44269, analyzed using the CheckM2 program, has a Completeness score of 100 with a Contamination score of –12.04, which casts doubt on its quality. Nevertheless, when using the Genomes Validator program, the genome size and GC content indicate the clear presence of extrachromosomal elements that affect the Contamination score, rather than contamination with foreign DNA, as evidenced by the research of the strain authors [16].

File	Species	Quality	Length	N50	GC	Reason
17892_1..NZ_JAJPNI010000010	P. aeruginosa	good	6 629 247	509 550	66,4	
178967_1..NZ_JAJPNH010000010	P. aeruginosa	bad	6 524 386	749 341	66,4	exsA not found
17896_7_2..NZ_JAJPKU01000001(	P. aeruginosa	bad	6 950 057	1 006 751	66,4	exsA not found
17897..NZ_JAJPNG010000010	P. aeruginosa	good	6 400 979	391 977	66,4	
17898_1..NZ_JAJPNF010000010	P. aeruginosa	good	6 432 087	511 042	66,4	
212_1..NZ_JAJPLU010000100	P. aeruginosa	average	6 260 559	63 774	66,8	sigX: 109/165 AK
212_2..NZ_JAJPLT010000010	P. aeruginosa	good	6 493 778	783 066	66,4	
215_4..NZ_JAJPLS010000010	P. aeruginosa	good	6 582 214	298 991	66,2	
220_2..NZ_JAJPLR010000010	P. aeruginosa	good	6 298 665	487 435	66,4	
224_1..NZ_JAJPLQ010000010	P. aeruginosa	average	6 523 464	322 858	66,2	AlgU: 47/193 AK
225_1..NZ_JAJPLP010000010	P. aeruginosa	average	6 419 808	377 166	66,3	endA: 0/237 AK
99_1..NZ_JAJPMG010000010	P. aeruginosa	good	6 829 566	675 464	66,4	
99_2..NZ_JAJPMF010000010	P. aeruginosa	good	6 426 869	414 815	66,4	
CriePir106..NZ_JAHYBC01000010(	P. aeruginosa	good	6 812 483	90 989	66,1	
CriePir111..NZ_JAHYBB01000010(	P. aeruginosa	good	6 951 545	78 424	65,7	
CriePir156..NZ_JAHYAV01000100	P. aeruginosa	average	6 689 553	10 010	65,6	dnaQ: 141/246 AK, endA: 0/237 AK, holB: 207/328 AK, tonB1: 215/342 AK
CriePir161..NZ_JAHYAU01000010	P. aeruginosa	average	6 800 283	23 819	65,9	tonB1: 215/342 AK
CriePir166..NZ_JAHYAT01000010(	P. aeruginosa	average	6 655 849	25 317	65,7	tonB1: 236/342 AK
CriePir178..NZ_JAHYAP01000010	P. aeruginosa	good	7 041 578	29 410	65,7	
CriePir191..NZ_JAHYAO01000010	P. aeruginosa	average	6 846 650	29 382	65,9	endA: 0/237 AK
CriePir198..NZ_JAHYAN01000010	P. aeruginosa	bad	6 374 609	45 212	66,3	dinB not found
CriePir199..NZ_JAHYAM0100001(	P. aeruginosa	good	6 838 465	38 717	66,0	
CriePir201..NZ_JAHYAL01000010(	P. aeruginosa	bad	6 634 250	36 597	65,8	sigX not found, exsA not found
P.aerug_8610	P. aeruginosa	good	6 924 285	246 611	65,6	
P.aerug_8612	P. aeruginosa	good	7 189 749	203 787	65,6	
P.aerug_8618	P. aeruginosa	good	7 137 324	160 374	64,9	
P.aerug_8633	P. aeruginosa	good	6 859 103	68 010	65,9	
Ps-agn-2308	P. aeruginosa	good	6 416 707	232 338	66,3	
Ps-agn-2350	P. aeruginosa	good	6 376 962	124 891	66,4	
Ps-agn-2424	P. aeruginosa	bad	8 752 214	9 282	64,2	Bad genome size
Ps-agn-2630	P. aeruginosa	average	6 799 119	16 295	66,3	
Ps-agn-2632	P. aeruginosa	average	6 763 523	19 158	66,3	
Ps-agn-2633	P. aeruginosa	average	6 719 361	13 924	66,3	
Ps-agn-2679	P. aeruginosa	good	6 640 636	32 082	66,3	
Ps-agn-2889	P. aeruginosa	bad	10 021 928	80 678	61,7	GC 61,7/66,0, Bad genome size
Ps-agn-2911	P. aeruginosa	bad	7 220 888	7 168	66,3	rpoH: 4/284 AK
Ps-agn-2935	P. aeruginosa	good	6 521 883	135 034	66,3	
Ps-agn-3458	P. aeruginosa	good	6 564 683	80 015	66,2	
Ps-agn-3835	P. aeruginosa	good	6 401 529	75 569	66,4	
Ps-agn-3842	P. aeruginosa	good	6 560 656	205 289	66,2	
SCPM-O-B-9017_(B-75_14)..NZ_J	P. aeruginosa	average	6 984 402	109 723	66,0	endA: 0/237 AK, cspD : 55/90 AK

Fig. 1. Practical demonstration of the program Genomes Validator; the results of genome analysis are presented in table format.

Discussion

The *oprI* gene was chosen as the species-defining gene for several reasons: its nucleotide sequence is 253 bp long, which allows for species identification even with very poor WGS data quality; the OprI protein plays an important role in binding to peptidoglycan, participates in immunological reactions, and is responsible for susceptibility to antimicrobial peptides [17–19]. This gene was chosen because one meta-analysis showed that it is successfully used to identify the *P. aeruginosa* species with high accuracy [20].

Housekeeping genes (*fur, algU, dinB, dnaQ, holA, holB, PA0472, fpvI, tonB1, cntL, sigX, capB, cspD, groES, rpoH*) were selected for validation of the chosen whole-genome sequences of *P. aeruginosa* based on their functional significance, as determined by a literature data analysis.

The *fur* gene is the main regulator of iron uptake in prokaryotic organisms, is essential for *P. aeruginosa* to cause pathogenesis, and for survival under iron-deficient conditions [21].

The sigma factor *algU* is a key stress response regulator that controls the expression of over 300 genes, plays a crucial role in virulence factor synthesis and pathogenesis thru quorum sensing, and enhances alginate production by increasing the expression of the *algD* operon [22].

SOS-mediated mutagenesis involves the products of the *dinB* gene, which perform translesion DNA synthesis, TLS (through damage), exhibiting low accuracy but helping to rapidly replicate DNA in response to various damaging agents. However, mutations accumulate, which in turn help acquire adaptive mechanisms in response to antibacterial drugs [23].

The DNA polymerase III ε subunit, encoded by the *dnaQ* gene, is very important and provides 3'-5' exonuclease activity, correcting mismatches encountered during DNA repair, which allows for the removal and correction of mismatched base pairs. Mutations in the *dnaQ* gene can disrupt these processes, leading to more than 1000-fold increase in the mutation rate in the genome [24]. The DNA polymerase III holoenzyme consists of δ and δ' subunits, which are encoded by the *holA* and *holB* genes, forming a complex with the ε subunit of the *dnaQ* gene and jointly participating in DNA repair [25]. The PA0472 gene encodes the RNA polymerase σ factor. It's difficult to judge what role a specific σ factor plays in the *P. aeruginosa* genome, but it is known that RNA polymerase σ factors perform a huge range of vital functions: promoter recognition, double-stranded DNA unwinding, binding to RNA polymerase, and transcription control. They are also involved in the transcription of specific regulons associated with the response to environmental changes and are included in iron transport [26].

One of the RNA polymerase σ-factors involved in iron assimilation processes is the FpvI protein, encoded by the *fpvI* gene, which is involved in regulating the uptake of the high-affinity siderophore pyoverdine, an important virulence factor as it can displace iron from the iron–transferrin complex [27].

P. aeruginosa has 3 genes in its genome that encode TonB proteins (*tonB1*, *tonB2*, and *tonB3*), and only the TonB1 protein, encoded by *tonB1*, interacts with TonB-dependent transporters involved in iron or heme uptake [28].

In addition to the main siderophores, *P. aeruginosa* produces another metallophore encoded by the *cntL* gene, called pseudopalin, which is essential for the uptake and utilization of zinc, cobalt, and nickel in its pathogenesis. Urease, which is a nickel-dependent enzyme, is produced by *P. aeruginosa*, while cobalt is essential for the cobalamin-dependent ribonucleotide reductase (NrdJab), which functions in biofilm formation under oxygen-limited conditions [29].

It is known that in *P. aeruginosa*, *sigX* is involved only in the transcription of its own gene and is largely responsible for the transcription of *oprF*, which encodes the major outer membrane protein OprF, which in turn is involved in several crucial functions: maintaining cell structure, outer membrane permeability, and recognition by the host immune system [30]. Deletion or knockout of the *algU* and *sigX* genes in the PAO1 genome disrupts biofilm formation [31].

The *capB* and *cspD* genes are responsible for encoding cold shock proteins involved in adaptation to cold in the environment [32].

Comparison of the quality metrics of the CheckM2 and Genomes Validator programs

«Genomes validator»	«CheckM2»	
	completeness	contamination
Good/High	99.99 ± 0.001	0.98 ± 0.203
Average	83.89 ± 1.865	2.23 ± 0.222
Poor/Low	81.01 ± 6.667	8.72 ± 3.708

Note. The confidence interval is indicated at $p < 0.05$

	CheckM2				Genomes Validator					
Strain	Completeness	Contamination	Contig_N50	GC_Content	Species	Quality	Length	N50	GC	Reason
294_2..JAJPNW010001000	100	1.06	191 133	0.64	P. aeruginosa	good	7672154	191133	63.8	
3392MAR21..JBKEPE0100001	100	0.07	252 389	0.66	P. aeruginosa	good	6585805	252389	66.3	
44269..JAGGDG010000986	100	12.04	225 018	0.66	P. aeruginosa	good	7829472	225018	66.3	
99_1..JAJPMG010000010	100	4.37	675 464	0.66	P. aeruginosa	good	6829566	675464	66.4	
Ps-agn-2889	100	35.65	80 678	0.62	P. aeruginosa	bad	10021928	80678	61.7	GC 61,7/66,0, Bad genome size

Fig. 2. Fragment of a table comparing the performance characteristics of the CheckM2 and Genomes Validator programs.

DOI: https://doi.org/10.36233/0372-9311-704

The *groES* gene encodes a heat shock protein that helps the microorganism survive at 42°C [33]. As is known, these resistance mechanisms are an integral part of the physiology of *P. aeruginosa* cells [33].

The σ^{32} factor, encoded by the *rpoH* gene, is the main regulator of the heat shock response, controlling the function of *groES*, among others [34].

The selected housekeeping genes confirmed their relevance in terms of their key role in the viability of *P. aeruginosa*, demonstrating the importance of their functioning for the biological processes of this microorganism. Furthermore, gene identification is not only based on the nucleotide sequence but is also translated into an amino acid sequence. This method was chosen to detect the stop codon in the gene and demonstrate not only its location in the genome but also its functionality. Thus, when assessing data quality, the absence of one or more of the selected genes will be considered a criterion for poor genome quality. If the N50 values for the genome selected for analysis are 10,000 or higher, but one of the candidate genes has a stop codon, it can be classified as a medium-quality genome. However, in our opinion, using a genome of this quality for phylogenetic analysis, SNP typing, or MLST analysis is not recommended. At the same time, searching for certain genes in the genome is possible, but without using them for typing the analyzed strain.

Despite the high N50 score and other evaluation parameters, the absence of two or more genes indicates poor WGS data quality. The N50 parameter is used to assess and compare the quality of genome assembly, allowing for the selection of the best among good/high-quality options.

The next criterion used to assess the quality of the WGS data was the GC content of the *P. aeruginosa* genomes. Analyzing the genomes of the strains using the CheckM2 program, we observed that genomes with Completeness scores > 97% and Contamination < 3% have a GC content ranging from 63.8% to 66.6%, which led us to establish threshold values of 65.2 ± 2.5%. The range was chosen wider to account for possible changes in the genomic composition. This criterion was supported by a literature review, which did not contradict our results and allowed us to include this parameter in a comprehensive quality assessment criterion for genome assemblies [35, 36]. This criterion demonstrates whether there is contamination of foreign DNA or reads from related species in the selected genome(s) for subsequent analysis.

WGS data validation using this criterion works as follows: if the genome selected for analysis falls within the established GC content values, it is considered a high-quality genome. If the selected genome does not fall within the established GC content values, it is considered a low-quality genome.

The quality of the genome is also assessed by the size of the *P. aeruginosa* whole-genome sequence. Thus, to evaluate the WGS data, the criteria for the minimum and maximum permissible genome sizes were used. The minimum genome size was 5.84 Mb, and the maximum was 8.26 Mb. The decision to use these values was based on literature data: for example, studies have reported that the auxiliary genome can vary within the range of 6.9–18.0% [38, 39]. The standard value for the length of the *P. aeruginosa* whole-genome sequence was taken as 6–7 Mb [37, 39].

Based on the above, the criterion for assessing the good quality of the *P. aeruginosa* genome will be a genome ranging in size from 5.84 to 8.26 Mb. If the analyzed genome falls outside the specified values, its quality will be assessed as poor or average, or the option of a more detailed and thorough analysis of this genome should be considered to exclude its structural features.

Genomes with an average level can be used limitedly for phylogenetic analysis, SNP typing, or MLST analysis, but they can be used to search for specific genes (without typing them) or for INDEL analysis.

Genomes with a low quality level are recommended not to be included in bioinformatics analysis and should be corrected by re-sequencing.

In addition to the CheckM2 and QUAST programs, which were selected as comparison tools, there are the SQUAT and Plantagora programs, but they do not meet the criteria of our research objects, as they are primarily developed for eukaryotic organisms. At the same time, CheckM2 is a tool developed for assessing the quality of prokaryotic genomes, while QUAST is a universal program. In developing our evaluation criteria, we tried to move away from complex tables with mathematical parameters assessing the quality of the WGS data provided by QUAST after the analysis. This involves the participation of bioinformatics specialists in the analysis and, in our opinion, does not fully reflect the quality of the WGS data, but rather assesses how well the genome assembly was performed [4]. At the same time, CheckM2 provides digital data on the parameters of the analyzed genome across various metrics, without drawing clear conclusions about the quality of the genome or whether it can be used for further research. The Contamination index does not always reflect the quality of the genomes of clinical isolates containing extrachromosomal elements.

Thus, we have tried, on the one hand, to select clear and concise parameters for evaluating WGS data, and on the other hand, to simplify the process for the user to obtain a specific result without resorting to in-depth bioinformatics analysis or using command linesк.

Conclusion

A comprehensive study was conducted in which we selected housekeeping genes that allow us to assess the quality of the *P. aeruginosa* WGS data. Quality assessment criteria have been defined: genome length

and GC content, which allow for the evaluation of the *P. aeruginosa* genome assembly.

Based on validated assessment criteria tested on a sample of genomes, the assembly of the *P. aeruginosa* genome can be classified into three categories based on the quality level of the source material: good, medium, and low. Good quality — the genome length is within the average genome size for the species ± 18%, the GC content is ± 2.5% of the average for *P. aeruginosa*, all essential genes have been found, and their protein product translation is not disrupted by a stop codon. Average quality – all essential genes found, but errors in their translation were detected due to the formation of a stop codon as a result of a sequencing error. Low quality — at least one gene in the life support system is missing, or the genome size or GC content does not match the value characteristic of the species.

An algorithm and a publicly available program for rapid analysis based on WGS data of *P. aeruginosa*, Genomes Validator, have been developed.

СПИСОК ИСТОЧНИКОВ | REFERENCE

1. Yang L.A., Chang Y.J., Chen S.H., et al. SQUAT: a Sequencing Quality Assessment Tool for data quality assessments of genome assemblies. *BMC Genomics*. 2019;19(Suppl. 9):238. DOI: https://doi.org/10.1186/s12864-019-5445-3
2. Barthelson R., McFarlin A.J., Rounsley S.D., Young S. Plantagora: modeling whole genome sequencing and assembly of plant genomes. *PLoS One*. 2011;6(12):e28436. DOI: https://doi.org/10.1371/journal.pone.0028436
3. Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072–5. DOI: https://doi.org/10.1093/bioinformatics/btt086
4. Parks D.H., Imelfort M., Skennerton C.T., et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25(7):1043–55. DOI: https://doi.org/10.1101/gr.186072.114
5. Chklovski A., Parks D.H., Woodcroft B.J., Tyson G.W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods*. 2023;20(8):1203–12. DOI: https://doi.org/10.1038/s41592-023-01940-w
6. Manchanda N., Portwood J.L. 2nd., Woodhouse M.R., et al. GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics*. 2020;21(1):193. DOI: https://doi.org/10.1186/s12864-020-6568-2
7. Manni M., Berkeley M.R., Seppey M., Zdobnov E.M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc*. 2021;1(12):e323. DOI: https://doi.org/10.1002/cpz1.323
8. Дятлов И.А., Миронов А.Ю., Шепелин А.П., Алешкин В.А. Состояние и тенденция развития клинической и санитарной микробиологии в Российской Федерации и проблема импортозамещения. *Клиническая лабораторная диагностика*. 2015;60(8):61–5. Dyatlov I.A., Mironov A.Yu., Shepelin A.P., Aleshkin V.A. The condition and tendencies of development of clinical and sanitary microbiology in the Russian Federation and problem of import substitution. *Russian Clinical Laboratory Diagnostics*. 2015;60(8):61–5. EDN: https://elibrary.ru/uiqjoh
9. Bolger A.M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–20. DOI: https://doi.org/10.1093/bioinformatics/btu170
10. Song L., Florea L., Langmead B. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol*. 2014;15(11):509. DOI: https://doi.org/10.1186/s13059-014-0509-9
11. 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(5):455–77. DOI: https://doi.org/10.1089/cmb.2012.0021
12. Wood D.E., Lu J., Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20(1):257. DOI: https://doi.org/10.1186/s13059-019-1891-0
13. Stover C.K., Pham X.Q., Erwin A.L., et al. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature*. 2000;406(6799):959–64. DOI: https://doi.org/10.1038/35023079
14. Носков А.К., Попова, А.Ю., Водопьянов, А.С., и др. Молекулярно-генетический анализ возбудителей бактериальных пневмоний, ассоциированных с COVID-19, в стационарах г. Ростова-на-Дону. *Здоровье населения и среда обитания*. 2021;(12):64–71. Noskov A.K., Popova A.Yu., Vodop'ianov A.S., et al. Molecular genetic analysis of the causative agents of COVID-19-associated bacterial pneumonia in hospitals of Rostov-on-Don. *Popul. Health Life Environ*. 2021;(12):64–71. DOI: https://doi.org/10.35627/2219-5238/2021-29-12-64-71 EDN: https://elibrary.ru/srsnhc
15. Wessel A.K., Liew J., Kwon T., et al. Role of *Pseudomonas aeruginosa* peptidoglycan-associated outer membrane proteins in vesicle formation. *J. Bacteriol*. 2013;195(2):213–9. DOI: https://doi.org/10.1128/JB.01253-12
16. Lu S., Chen K., Song K., et al. Systems serology in cystic fibrosis: anti-pseudomonas IgG1 responses and reduced lung function. *Cell Rep. Med*. 2023;4(10):101210. DOI: https://doi.org/10.1016/j.xcrm.2023.101210
17. Sabzehali F., Goudarzi H., Chirani A.S., et al. Development of multi-epitope subunit vaccine against *Pseudomonas aeruginosa* using OprF/OprI and PopB proteins. *Arch. Clin. Infect. Dis*. 2021;16(4). DOI: https://doi.org/10.22038/IJBMS.2022.61448.13595
18. Tang Y., Ali Z., Zou J., et al. Detection methods for *Pseudomonas aeruginosa:* history and future perspective. *Rsc Advances*. 2017;7(82):51789–800. DOI: https://doi.org/10.1039/c7ra09064a
19. Sevilla E., Bes M.T., Peleato M.L., Fillat M.F. Fur-like proteins: Beyond the ferric uptake regulator (Fur) paralog. *Arch. Biochem. Biophys*. 2021;701:108770. DOI: https://doi.org/10.1016/j.abb.2021.108770
20. Kar A., Mukherjee S.K., Hossain S.T. Regulatory role of PA3299.1 small RNA in *Pseudomonas aeruginosa* biofilm formation via modulation of algU and mucA expression. *Biochem. Biophys. Res. Commun*. 2025;748:151348. DOI: https://doi.org/10.1016/j.bbrc.2025.151348
21. Fahey D., O'Brien J., Pagnon J., et al. DinB (DNA polymerase IV), ImuBC and RpoS contribute to the generation of ciprofloxacin-resistance mutations in *Pseudomonas aeruginosa. Mutat. Res*. 2023;827:111836. DOI: https://doi.org/10.1016/j.mrfmmm.2023.111836
22. Dekker J.P. Within-host evolution of bacterial pathogens in acute and chronic infection. *Annu. Rev. Pathol*. 2024; 19:203–26. DOI: https://doi.org/10.1146/annurev-pathmechdis-051122-111408
23. Spinnato M.C., Lo Sciuto A., Mercolino J., et al. Effect of a defective clamp loader complex of DNA polymerase III on growth and SOS response in *Pseudomonas aeruginosa*. *Microorganisms*. 2022;10(2):423. DOI: https://doi.org/10.3390/microorganisms10020423

24. Potvin E., Sanschagrin F., Levesque R.C. Sigma factors in *Pseudomonas aeruginosa*. *FEMS Microbiol. Rev*. 2008;32(1):38–55. DOI: https://doi.org/10.1111/j.1574-6976.2007.00092.x
25. Cornelis P., Tahrioui A., Lesouhaitier O., et al. High affinity iron uptake by pyoverdine in *Pseudomonas aeruginosa* involves multiple regulators besides Fur, PvdS, and FpvI. *Biometals*. 2023;36(2):255–61. DOI: https://doi.org/10.1007/s10534-022-00369-6
26. Peukert C., Gasser V., Orth T., et al. Trojan horse siderophore conjugates induce *Pseudomonas aeruginosa* suicide and qualify the TonB protein as a novel antibiotic target. *J. Med. Chem*. 2023;66(1):553–76. DOI: https://doi.org/10.1021/acs.jmedchem.2c01489
27. Ghssein G., Ezzeddine Z. A Review of *Pseudomonas aeruginosa* metallophores: pyoverdine, pyochelin and pseudopaline. *Biology (Basel)*. 2022;11(12):1711. DOI: https://doi.org/10.3390/biology11121711
28. Duchesne R., Bouffartigues E., Oxaran V., et al. A proteomic approach of SigX function in *Pseudomonas aeruginosa* outer membrane composition. *J. Proteomics*. 2013;94:451–9. DOI: https://doi.org/10.1016/j.jprot.2013.10.022
29. Østergaard M.Z., Nielsen F.D., Meinfeldt M.H., Kirkpatrick C.L. The uncharacterized PA3040-3042 operon is part of the cell envelope stress response and a tobramycin resistance determinant in a clinical isolate of *Pseudomonas aeruginosa. Microbiol. Spectr*. 2024;12(8):e0387523. DOI: https://doi.org/10.1128/spectrum.03875-23
30. Li S., Weng Y., Li X., et al. Acetylation of the CspA family protein CspC controls the type III secretion system through translational regulation of exsA in *Pseudomonas aeruginosa. Nucleic Acids Res*. 2021;49(12):6756–70. DOI: https://doi.org/10.1093/nar/gkab506
31. Williamson K.S., Dlakić M., Akiyama T., Franklin M.J. The *Pseudomonas aeruginosa* RpoH (σ32) regulon and its role in essential cellular functions, starvation survival, and antibiotic tolerance. *Int. J. Mol. Sci*. 2023;24(2):1513. DOI: https://doi.org/10.3390/ijms24021513
32. LaBauve A.E., Wargo M.J. Growth and laboratory maintenance of *Pseudomonas aeruginosa. Curr. Protoc. Microbiol*. 2012;Chapter 6:6E.1. DOI: https://doi.org/10.1002/9780471729259.mc06e01s25
33. Li Y., Bhagirath A., Badr S., et al. The Fem cell-surface signaling system is regulated by ExsA in *Pseudomonas aeruginosa* and affects pathogenicity. *iScience.* 2024;28(1):111629. DOI: https://doi.org/10.1016/j.isci.2024.111629
34. Valot B., Guyeux C., Rolland J.Y., et al. What it takes to be a *Pseudomonas aeruginosa*? The core genome of the opportunistic pathogen updated. *PLoS One*. 2015;10(5):e0126468. DOI: https://doi.org/10.1371/journal.pone.0126468
35. Subedi D., Kohli G.S., Vijay A.K., et al. Accessory genome of the multi-drug resistant ocular isolate of *Pseudomonas aeruginosa* PA34. *PLoS One*. 2019;14(4):e0215038. DOI: https://doi.org/https://doi.org/10.1371/journal.pone.0215038
36. Ozer E.A., Allen J.P., Hauser A.R. Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. *BMC Genomics*. 2014;15(1):737. DOI: https://doi.org/10.1186/1471-2164-15-737
37. Dettman J.R., Rodrigue N., Aaron S.D., Kassen R. Evolutionary genomics of epidemic and nonepidemic strains of *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA*. 2013;110(52):21065–70. DOI: https://doi.org/10.1073/pnas.1307862110

Information about the authors

Alexey A. Kovalevich✉ — junior researcher, Laboratory of molecular biology of natural focal and zoonotic infections, Rostov-on-Don Antiplague Scientific Research Institute, Rostov-on-Don, Russia, kovalevich_aa@antiplague.ru, https://orcid.org/0000-0001-6926-0239

Ruslan V. Pisanov — Cand. Sci. (Biol.), leading researcher, Laboratory of molecular biology of natural focal and zoonotic infections, Rostov-on-Don Antiplague Scientific Research Institute, Rostov-on-Don, Russia, pisanov.ruslan@yandex.ru, https://orcid.org/0000-0002-7178-8021

Alexey S. Vodopianov — Cand. Sci. (Med.), leading researcher, Laboratory of molecular biology of natural focal and zoonotic infections, Rostov-on-Don Antiplague Scientific Research Institute, Rostov-on-Don, Russia, vodopyanov_as@antiplague.ru, https://orcid.org/0000-0002-9056-3231

Authors' contribution: *Kovalevich A.A.* — research concept and design, data analysis, writing; *Vodopyanov A.S.* — software development, debugging, editing, research concept; *Pisanov R.V.* — general research guidance, attract financing, drafting the work, final approval of the version for publication. All authors made a substantial contribution to the conception of the work, acquisition, analysis, interpretation of data for the work, drafting and revising the work, final approval of the version to be published.

The article was submitted 28.07.2025;
accepted for publication 29.09.2025;
published 31.10.2025

Информация о авторах

Ковалевич Алексей Александрович✉ — м. н. с. лаб. молекулярной биологии природно-очаговых и зоонозных инфекций Ростовского-на-Дону научно-исследовательского противочумного института, Ростов-на-Дону, Россия, kovalevich_aa@antiplague.ru, https://orcid.org/0000-0001-6926-0239

Писанов Руслан Вячеславович — канд. биол. наук, в. н. с., зав. лаб. молекулярной биологии природно-очаговых и зоонозных инфекций Ростовского-на-Дону научно-исследовательского противочумного института, Ростов-на-Дону, Россия, pisanov.ruslan@yandex.ru, https://orcid.org/0000-0002-7178-8021

Водопьянов Алексей Сергеевич — канд. мед. наук, в. н. с. лаб. молекулярной биологии природно-очаговых и зоонозных инфекций Ростовского-на-Дону научно-исследовательского противочумного института, Ростов-на-Дону, Россия, vodopyanov_as@antiplague.ru, https://orcid.org/0000-0002-9056-3231

Участие авторов: *Ковалевич А.А.* — концепция и дизайн исследования, анализ данных, написание текста; *Водопьянов А.С.* — разработка программного обеспечения, отладка, редактирование, концепция исследования; *Писанов Р.В.* — общее руководство исследования, рецензирование и научное редактирование текста рукописи, окончательное утверждение версии для публикации. Все авторы внесли существенный вклад в проведение поисково-аналитической работы и подготовку статьи, прочли и одобрили финальную версию до публикации

Статья поступила в редакцию 28.07.2025;
принята к публикации 29.09.2025;
опубликована 31.10.2025