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

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).

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

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

Введение

С развитием технологий высокопроизводительного секвенирования и снижением их стоимости объёмы производимых данных о геномах растут в геометрической прогрессии. Проекты, использующие большие массивы данных полногеномного секвенирования (whole-genome sequencing, WGS), имеют множество преимуществ: повышается статистическая мощность, появляется возможность проверять различные гипотезы о микро- и макроэволюции геномов.

Постоянное совершенствование технологий секвенирования и биоинформатического анализа повысило значимость WGS в биологии, медицине, фармацевтике и сельском хозяйстве, стимулируя проведение сравнительных геномных исследований. Однако рост числа проектов и лабораторий, занимающихся секвенированием, привёл к увеличению количества сборок геномов, не всегда пригодных для анализа. Это подчеркнуло необходимость оценки качества данных полногеномных сборок для исследователей, использующих эти данные. В связи с этим возникла потребность в разработке стандартных показателей для сравнения качества сборок и аннотаций геномов, а также для оценки эффективности различных методов их получения.

Недавние исследования оценки качества сборки геномов были сосредоточены либо на контроле качества перед сборкой, либо на оценке сборки с точки зрения непрерывности и правильности. Однако оценка корректности зависит от эталона и неприменима для проектов сборки *de novo*. Следовательно, стоит изучить методы, позволяющие получать отчёты об оценке качества как после сборки, так и до неё, для проверки качества/корректности сборки *de novo* и входных данных [1].

Для сборок генома такие показатели, как количество контигов, количество скафолдов, N50 (максимальная длина контига, при котором суммарная длина всех контигов, не короче этой величины, составляет не менее 50% от общей длины всей контигов в сборке), дают только краткое представление о качестве генома, не всегда отражая его аналитическую пригодность.

В свою очередь на сегодняшний день существует достаточное количество ресурсов и инструментов для постаналитического этапа работы, а также оценки качества геномов: Picard[1], SQUAT [1], Plantagora [2], QUAST [3], CheckM1 [4], CheckM2 [5], GenomeQC [6], BUSCO [7]. Однако они являются унифицированными и представляют собой алгоритмы разной направленности, иногда удобные для анализа только эукариотических организмов, при этом не учитывающие особенностей организации генома конкретного вида. Одним из наиболее универсальных и широко применяемых инструментов, использующих гены для оценки данных WGS, является BUSCO. В отличие от упомянутых выше решений, BUSCO фокусируется на анализе геномов, используя эволюционно консервативные гены-ортологи, которые считаются «универсальными» для некоторых таксономических групп (бактерий, грибов, растений или животных). Однако BUSCO не даёт ответа о качестве анализируемого генома, выдавая только процент обнаруженных/необнаруженных генов-ортологов, и заключительный ответ должен сформировать сам специалист. Вместе с тем гены-ортологи могут утрачиваться без потери

[1] Pickard Tools. GitHub repository. Broad Institute; 2019. URL: http://broadinstitute.github.io/picard

жизнеспособности бактерий, в отличие от генов домашнего хозяйства, что может привести к ошибке оценки качества генома.

В настоящее время WGS возбудителей инфекционных заболеваний широко используется для их изучения, определения их происхождения и распространения. Для оценки качества такого большого количества данных необходимы отечественные программные инструменты. Последнее особенно важно ввиду того, что импортозамещение становится одной из стратегических задач в условиях, когда доступ к зарубежным технологиям, а также иностранным банкам данных затруднён [8].

Критериев оценки данных WGS для представителей *Pseudomonas aeruginosa* на сегодняшний день нет. Существуют программные сервисы, которые осуществляют оценку в общих (неспецифических) критериях (N50, количество контигов и т. п.) и не учитывают особенности конкретного микроорганизма.

Цель исследования — создание алгоритма и критериев для оценки качества данных WGS представителей вида *P. aeruginosa*, а также создание отечественного программного обеспечения, способного оценить качество данных WGS.

Материалы и методы

В работе использовали 108 геномов штаммов *P. aeruginosa*: 24 штамма получены из Коллекции патогенных микроорганизмов Ростовского-на-Дону противочумного института Роспотребнадзора (выделены в Ростове-на-Дону, Хабаровске, Мариуполе в 2022–2024 гг.), 84 штамма получены из международной базы NCBI. WGS проведено в ходе реализации федерального проекта социально-экономического развития Российской Федерации до 2030 г. «Санитарный щит страны — безопасность для здоровья (предупреждение, выявление, реагирование)». Секвенирование проводили на платформе «MiSeq» («Illumina») с использованием набора для секвенирования «MiSeq Reagent Kit v2 (500-cycles)» («Illumina»). Этот метод позволяет получать прочтения длиной 2 × 251 нуклеотидов, покрытие геномов составляло 8–20.

Оценку первичных данных секвенирования проводили с использованием программы FastQC. Собранные данные WGS анализировали с помощью программы QUAST[2] [4]. Для тримминга и коррекции ридов использовали алгоритмы Trimmomatic [9] и Lighter [10]. Сборку геномов, представленных в виде ридов, проводили с использованием программы Spades [11]. Все геномы прошли первичную оценку с помощью программы Kraken 2, позволяющей определять принадлежность фрагментов ДНК к различным прокариотическим видам [12]. В качестве референс-генома были использованы данные WGS штамма PAO1 из международной базы NCBI [13].

Авторское программное обеспечение разрабатывали на языках программирования Java и Python. Алгоритм поиска последовательностей генов в сборке осуществляли с помощью локального выравнивания Смита–Ватермана с минимальным порогом сходства 80%.

Рассчитывали доверительный интервал (формат представления данных — $M \pm SD$), различия считали значимыми при $p < 0{,}05$.

Результаты

Известно, что в составе генома возбудителя синегнойной инфекции имеется ряд генов, критически важных для его жизнедеятельности. Такие гены получили название «гены жизнеобеспечения» (housekeeping genes). Очевидно, что если в данных WGS отсутствует какой-либо из этих генов, то это является ошибкой секвенирования и/или сборки генома. Именно данная особенность и была положена в основу предлагаемого нами алгоритма — в сиквенсе хорошего качества должны обнаруживаться все гены жизнеобеспечения. Разумеется, при этом большое значение имеет подбор генов, по которым будет проводиться контроль качества.

Один из критериев для оценки качества геномов, который был продуман нами для алгоритма, — это выбор генов, которые будут отобраны по следующим критериям:

- нуклеотидные последовательности в пределах 1000 пар оснований (п. о.);
- ген должен быть однокопийным;
- ген должен непосредственно участвовать в физиологической деятельности микроорганизма, выполняя основные функции для жизнедеятельности;
- ген должен присутствовать у всех штаммов *P. aeruginosa*.

Для оперативной видовой идентификации *P. aeruginosa* был выбран ген *oprI.* Основной задачей является оценка качества данных секвенирования не только на основе идентификации генов жизнеобеспечения, но и на транслировании их последовательностей. Учитывая, что данные гены являются критически важными для существования микробной клетки, их отсутствие в геноме или критические ошибки трансляции (стоп-кодоны) расцениваются как ошибка секвенирования.

Гены жизнеобеспечения, по которым осуществлялась валидация выбранных полногеномных последовательностей *P. aeruginosa*: *fur*, *algU*, *dinB*, *dnaQ*, *holA*, *holB, PA0472*, *fpvI*, *tonB1*, *cntL*, *sigX*, *capB*, *cspD*, *groES*, *rpoH.* Выбранные гены являются обязательными для функционирования и выживания в окружающей среде и макроорганизме. В каче-

[2] Andrews S. FastQC: a quality control tool for high throughput sequence data; 2010. URL: https://bioinformatics.babraham.ac.uk/projects/fastqc/

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

стве критериев оценки полногеномных последовательностей были выбраны следующие параметры: GC-состав геномов *P. aeruginosa*, размер полногеномной последовательности *P. aeruginosa*, значение скафолда N50.

После проведения исследования и выбора критериев оценки качества геномов разработано программное обеспечение «Genomes Validator», которое для удобства предполагает работу в «пакетном режиме», анализирует неограниченное количество геномов, результаты выдаются в табличном виде. При этом для каждого генома указываются название исходного файла, вид, качество (bad, average, good), длина, показатель N50, показатель GC-состава, причина невалидности генома (**рис. 1**).

Разработанная программа «Genomes Validator» является кросс-платформенной, имеет графический интерфейс, не требует установки, позволяет анализировать несколько геномов сразу и доступна для скачивания на сайте https://github.com/alexeyvod/GenomesValidator, обладает интуитивно понятным интерфейсом и удобна для пользователей без навыков программирования.

Валидация программы была проведена на выборке из 108 полногеномных последовательностей штаммов *P. aeruginosa.* По итогам валидации были идентифицированы геномы хорошего (63%), среднего (29%), плохого (8%) качества. Из дальнейшего анализа были выведены 37% анализируемых геномов (среднего и плохого качества), что может помочь избежать ошибок при дальнейших расчётах с использованием биоинформатических методов. Средний показатель N50 среди выборки был равен 1 250 527.

Параметры N50, длина генома, GC-состав были идентичными показателям программ, взятых для сравнения: «CheckM2» и «QUAST». Однако эти

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

Рис. 1. Демонстрация работы программы «Genomes Validator», результат анализа геномов приводится в табличном формате.

программы не предоставляют параметры оценки качества геномов.

При оценке качества бактериальных геномов с использованием программного обеспечения «CheckM2» нами установлено, что геномы со значением Completeness, равным 100, демонстрировали значительную вариабельность показателя контаминации (Contamination). Вместе с тем применение инструмента «Genomes Validator» позволило дополнительно оценить размер полногеномной последовательности, что может быть более информативным для практического анализа данных WGS (**рис. 2**). Данный параметр позволяет предварительно судить о наличии внехромосомных элементов в геноме исследуемого штамма. Следует учитывать, что контаминация ДНК посторонней микрофлорой, как правило, отражается на общем GC-составе и существенном изменении размера генома, тогда как присутствие плазмид или других мобильных генетических элементов не вызывает существенных изменений этого параметра.

На основании проведённого статистического анализа параметров Completeness и Contamination (**таблица**) выявлено, что значения Contamination в диапазоне от 2 до 8 могут свидетельствовать о возможной низкой достоверности полученных данных WGS. Однако такие результаты также могут быть обусловлены специфическими особенностями генома клинического изолята. Так, геном штамма Ps-agn-2889, проанализированный в программе «CheckM2», имеет показатель Completeness = 100 при Contamination = –35,65, однако из полученных данных причина контаминации не ясна. При анализе в программе «Genomes Validator» увеличенный в 1,5 раза размер генома и GC-состав указывают на явную контаминацию посторонней бактериальной ДНК. Геном клинического штамма 44269 проанализированный в программе «CheckM2», имеет показатель Completeness = 100 при Contamination = –12,04, что ставит под сомнение его качество. Тем не менее при использовании программы «Genomes Validator» размер генома и GC-состав указывают на явное присутствие внехромосомных элементов, которые влияют на показатель Contamination, а не на контаминацию чужеродной ДНК, что следует из исследования авторов штамма [16].

Сравнение показателей качества программ «CheckM2» и «Genomes Validator»

«Genomes validator»	«CheckM2»	
	completeness	contamination
Good	99,99 ± 0,001	0,98 ± 0,203
Average	83,89 ± 1,865	2,23 ± 0,222
Bad	81,01 ± 6,667	8,72 ± 3,708

Примечание. Указаны данные при $p < 0,05$.

Обсуждение

В качестве гена, определяющего видовую принадлежность, был выбран *oprI* по ряду причин: нуклеотидная последовательность 253 п. о., что даёт возможность определить видовую принадлежность даже в случае очень плохого качества данных WGS; белок OprI выполняет важную роль в связывании с пептидогликаном, участвует в иммунологических реакциях и чувствительности к антимикробным пептидам [15–17]. Этот ген был выбран, поскольку одно из метаисследований показало, что он успешно применяется для идентификации вида *P. aeruginosa* с высокой точностью [18].

Гены жизнеобеспечения, по которым будет осуществляться валидация выбранных полногеномных последовательностей *P. aeruginosa* (*fur*, *algU*, *dinB*, *dnaQ*, *holA*, *holB, PA0472*, *fpvI*, *tonB1*, *cntL*, *sigX*, *capB*, *cspD*, *groES*, *rpoH*), были выбраны в результате анализа данных литературных на основании их функциональной значимости.

Ген *fur* является основным регулятором поглощения железа у прокариотических организмов, необходим *P. aeruginosa* для реализации патогенеза, а также выживания в условиях дефицита железа [19].

Сигма-фактор *algU* — ключевой регулятор реакции на стресс, который контролирует экспрессию более 300 генов, играет важнейшую роль в синтезе факторов вирулентности и патогенезе посредством кворум-сенсинга, усиливает выработку альгината, повышая экспрессию оперона *algD* [20].

В SOS-опосредованный мутагенез вовлекаются продукты гена *dinB*, которые осуществляют трансслезионный синтез ДНК (через повреждение), демонстрируя низкую точность, но при этом помогая оперативно реплицировать ДНК в ответ на различные агенты, её повреждающие. Однако происходит накопление мутаций, которые в свою очередь помогают приобретать адаптивные механизмы в ответ на антибактериальные препараты [21].

	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

Рис. 2. Фрагмент таблицы сравнительной характеристики работы программ «CheckM2» и «Genomes validator».

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

ДНК-полимераза III ε-субъединица, кодируемая геном *dnaQ*, очень важна и обеспечивает 3′-5′-экзонуклеазную активность, корректируя несоответствия, встречающиеся при репарации ДНК, что позволяет удалять и исправлять несовпадающие пары оснований. Мутации гена *dnaQ* могут приводить к нарушению этих процессов, при этом частота мутаций в геноме увеличивается более чем в 1000 раз [22]. Холофермент ДНК-полимераза III состоит из δ-, δ′-субъединиц, которые кодируются генами *holA*, *holB*, образуя сложный комплекс с ε-субъединицей гена *dnaQ* и совместно участвуя в репарации ДНК [23]. Ген *PA0472* кодирует σ-фактор РНК-полимеразы. Тяжело судить, какую роль в геноме *P. aeruginosa* выполняет конкретный σ-фактор, но известно, что σ-факторы РНК-полимераз выполняют огромный спектр жизненно необходимых функций: распознавание промотора, расплетение двухцепочечной ДНК, связывание с РНК-полимеразой, контроль транскрипции. Они участвуют также в транскрипции специфических регулонов, связанных с ответом на изменения окружающей среды, включены в транспорт железа [24].

Одним из σ-факторов РНК-полимеразы, участвующей в процессах ассимиляции железа, является белок FpvI, кодируемый геном *fpvI*, который вовлечён в процессы регуляции поглощения высокоаффинного сидерофора — пиовердина, являющегося важным фактором вирулентности, т. к. способен вытеснять железо из комплекса железо–трансферрин [25].

P. aeruginosa имеет в своём геноме 3 гена, кодирующие белки TonB (*tonB1*, *tonB2* и *tonB3*), и только белок TonB1, кодируемый *tonB1,* взаимодействует с TonB-зависимыми переносчиками, задействованными в поглощении железа или гема [26].

Помимо основных сидерофоров, *P. aeruginosa* вырабатывает ещё один металлофор, кодируемый геном *cntL*, под названием псевдопалин, необходимый в усвоении и использовании в своём патогенезе цинка, кобальта и никеля. Уреаза, которая является никельзависимым ферментом, вырабатывается *P. aeruginosa,* в то время как кобальт необходим для кобаламинзависимой рибонуклеотидредуктазы (NrdJab), выполняющей функцию формирования биоплёнки в условиях ограниченного доступа кислорода [27].

Известно, что у *P. aeruginosa sigX* участвует только в транскрипции собственного гена и в значительной степени отвечает за транскрипцию *oprF*, кодирующего основной белок внешней мембраны OprF, который, в свою очередь, участвует в нескольких важнейших функциях: поддержание структуры клетки, проницаемость внешней мембраны, распознавание иммунной системы хозяина [28]. При удалении или нокаутированнии генов *algU* и *sigX* в геноме PAO1 нарушается формирование биоплёнки [29].

Гены *capB* и *cspD* отвечают за кодирование белков холодового шока, участвующих в адаптации к холоду в окружающей среде [30].

Ген *groES* кодирует белок теплового шока, который помогает выживать микроорганизму при температуре 42°C [31]. Как известно, эти механизмы устойчивости являются неотъемлемой частью физиологии клеток *P. aeruginosa* [31].

σ^{32}-Фактор, кодируемый геном *rpoH*, выступает основным регулятором реакции на тепловой шок, контролируя в том числе работу *groES* [32].

Выбранные гены жизнеобеспечения подтвердили свою релевантность с точки зрения их ключевой роли в жизнедеятельности *P. aeruginosa*, продемонстрировав важность их функционирования для биологических процессов данного микроорганизма. Кроме того, идентификация генов не только идёт по нуклеотидной последовательности, но и транслируется в аминокислотную. Этот способ был выбран для того, чтобы детектировать стоп-кодон в гене и демонстрировать не только его нахождение в геноме, но и функциональность. Таким образом, при оценке качества данных критерием плохого качества генома будет считаться отсутствие одного или более из выбранных генов. Если у генома, выбранного в анализ, показатели N50 будут от 10 000 и выше, но один из генов-кандидатов имеет стоп-кодон, то его можно отнести к геному среднего качества. Вместе с тем, по нашему мнению, использовать геном такого качества для филогенетического анализа, SNP-типирования или MLST-анализа не рекомендуется. При этом поиск в геноме тех или иных генов осуществлять можно, но без использования их для типирования анализируемого штамма.

Несмотря на высокие показатели N50, а также другие параметры оценки, отсутствие двух и более генов указывает на плохое качество данных WGS. Параметр N50 используется для оценки и сравнения качества сборки генома, позволяя выбирать лучшую среди вариантов хорошего/высокого качества (good).

Следующий критерий, с помощью которого оценивали качество данных WGS, стал GC-состав геномов *P. aeruginosa*. Анализируя геномы штаммов с помощью программы «CheckM2», мы заметили, что геномы с показателями Completeness > 97% и Contamination < 3% имеют GC-состав от 63,8 до 66,6%, в связи с чем нами были установлены пороговые значения 65,2 ± 2,5%. Диапазон был выбран шире, чтобы учесть возможные изменения геномного состава. Этот критерий был подкреплён анализом источников литературы, что не противоречило нашим результатам и позволило включить данный параметр в комплексный критерий оценки

качества геномных сборок [33, 34]. Указанный критерий демонстрирует, присутствует ли «контаминация» чужеродной ДНК или ридов родственных видов выбранного генома/геномов для последующего анализа.

Валидация данных WGS с использованием этого критерия происходит по следующему принципу: если выбранный для анализа геном попадает в установленные значения GC-состава, он оценивается как геном хорошего качества. В том случае, если выбранный геном не попадает в установленные значения GC-состава, он оценивается как геном плохого качества.

Качество генома также оценивается по размеру полногеномной последовательности *P. aeruginosa*. Так, для оценки данных WGS были использованы критерии минимального и максимального допустимых размеров генома. За минимальное значение генома был принят размер 5,84 Мб, за максимальное — 8,26 Мб. Основой для принятия решения использования этих значений послужили данные литературы: так, в исследованиях были озвучены показатели, при которых вспомогательный геном может варьировать в пределах 6,9–18,0% [35, 36]. За стандартное значение длины полногеномной последовательности *P. aeruginosa* был принят параметр 6–7 Мб [35, 37].

На основании вышеизложенного, критерием оценки хорошего качества генома *P. aeruginosa* будет являться геном размерами 5,84–8,26 Мб. В случае если анализируемый геном будет выходить за указанные значения, его качество будет оценено как плохое либо среднее, или же следует рассмотреть вариант более детального и пристального анализа данного генома, дабы исключить его структурные особенности.

Геномы, имеющие «средний» уровень, могут быть использованы ограниченно для филогенетического анализа, SNP-типирования или MLST-анализа, однако они могут быть использованы для поиска тех или иных генов (без их типирования) или INDEL-анализа.

Геномы, имеющие «плохой» уровень качества, рекомендуется не брать в биоинформационный анализ и корректировать путём проведения повторного секвенирования.

Кроме программ «CheckM2» и «QUAST», выбранных в качестве инструментов сравнения, существуют программы «SQUAT» и «Plantagora», однако они не соответствуют критериям наших объектов исследования, т. к. разработаны преимущественно для эукариотических организмов. В то же время «CheckM2» является инструментом, разработанным для оценки качества прокариотических геномов, а «QUAST» — универсальная программа. В разработке нашего критерия оценки мы постарались уйти от сложных таблиц с математическими параметрами, оценивающими качество данных WGS, которые предоставляет «QUAST» по итогам анализа. Это предполагает участие в анализе специалистов-биоинформатиков, а также, по нашему мнению, не отражает в полной степени качество WGS данных, а оценивает то, как качественно была проведена сборка генома [3]. В то же время «CheckM2» предоставляет цифровые данные о показателях анализируемого генома по различным параметрам, не давая чётких заключений о том, какого качества геном и можно ли его использовать для дальнейшей работы. Показатель Contamination в некоторых случаях не отражает качества геномов клинических изолятов, содержащих внехромосомные элементы.

Таким образом, мы постарались, с одной стороны, подобрать ясные и чёткие параметры оценки данных WGS, с другой стороны, упростить для пользователя получение конкретного результата, не прибегая к углублённому биоинформатическому анализу, без использования командных строк.

Заключение

Проведено комплексное исследование, в котором нами были отобраны гены жизнеобеспечения, позволяющие оценить качество данных WGS *P. aeruginosa*. Определены критерии оценки качества: размер длины генома, GC-состав, которые позволяют оценить сборку генома *P. aeruginosa*.

На основе валидированных критериев оценки, проверенных на выборке геномов, можно классифицировать сборку генома *P. aeruginosa* по уровню качества исходного материала на три категории: хорошее, среднее и низкое. Хорошее качество — длина генома соответствует среднему размеру генома для вида ± 18%, доля GC составляет ± 2,5% от среднего показателя *P. aeruginosa*, все гены жизнеобеспечения найдены, трансляция их белкового продукта не нарушена стоп-кодоном. Среднее качество — все гены жизнеобеспечения найдены, но обнаружены ошибки их трансляции вследствие образования стоп-кодона из-за ошибки секвенирования. Низкое качество — не найден хотя бы один ген системы жизнеобеспечения либо размер генома или GC-состав не соответствует значению, характерному для вида.

Разработаны алгоритм и общедоступная программа для оперативного анализа на основе данных WGS *P. aeruginosa* «Genomes Validator».

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

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

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

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

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

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

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

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

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