Math-Net.Ru
Общероссийский математический портал

V. V. Panyukov, S. S. Kiselev, O. V. Alikina, N. N. Nazipova, O. N. Ozoline, Короткие уникальные последовательности в бактериальных геномах как штамм- и видоспецифические маркеры, *Матем. биология и биоинформ.*, 2017, том 12, выпуск 2, 547–558

DOI: 10.17537/2017.12.547

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением
http://www.mathnet.ru/rus/agreement

Параметры загрузки:
IP: 3.12.136.186
5 октября 2024 г., 08:18:16

Mathematical Biology and Bioinformatics
2017. V. 12. № 2. P. 547-558. doi: 10.17537/2017.12.547

========== **BIOINFORMATICS** ==========

UDC: 579:252

Short unique sequences in bacterial genomes as strain- and species-specific signatures

©2017 Panyukov V.V.[1,2], Kiselev S.S.[3], Alikina O.V.[3], Nazipova N.N.[1,2], Ozoline O.N.[*1,3]

[1]*142290, Pushchino Research Center of Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation*
[2]*142290, Institute of Mathematical Problems of Biology – the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation*
[3]*142290, Institute of Cell Biophysics of Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation*

Abstract. The paper presents a new approach for phylotyping that can be potentially used for pure cultures and for mixed bacterial populations. It is based on the use of short unique nucleotide sequences (*k*-mers) that are present in the genomes of all strains of the same species and are absent in bacterial genomes of other taxonomic groups. We show that the number *N* of such sequences depends on the percentage bias towards A/T or G/C base pairs, increasing for genomes with approximately equal composition. We found that the largest contribution to the set of primarily unique sequences is given by 16–17-mers, while sigmoidal curves reflecting the dependence of *N* on the length of *k*-mers showed the maximum slope increment ($\Delta N/\Delta k$) for $k = 17, 18$. Unique sequences of the length 16–18 bases can therefore be offered as potential markers. Comparing the sets of unique *k*-mers in the genomes of four *Enterobacter* strains, we estimated the level of their intraspecies stability and interspecies plasticity. As a result, we suggest discriminatory subsets as stencils for phylotyping, thereby increasing the list of genotyping markers with signatures of the new type.

Key words: microbiomes, bacterial genomes, genotyping, unique nucleotide sequences.

INTRODUCTION

The danger of pathogenic microorganisms has always been an important factor that stimulated the development of methods for their reliable identification. The last two decades were marked by a particularly sharp progress in this field, which largely supplanted the phenotypic methods of classification (morphology of cells and colonies, cultivation conditions, resistance to antibiotics, serotype, toxicity, pathogenicity, etc.), replacing them with more precise genotypic approaches [1, 2]. The currently used methods of genotyping are very diverse. Some of them do not need any information about the nucleotide sequence of the genomes being analyzed. Thus, in the Pulsed-Field Gel Electrophoresis method (PFGE), only specific restriction with rare-cutting endonucleases and electrophoretic fractionation of long fragments (5000–1000000 bp) are required [3, 4]. As a result, an easily interpreted distribution of the restriction fragments on gels is obtained that makes it possible to distinguish one strain from another or to conclude that they are identical. In RAPD method (Random Amplified Polymorphic DNA), a polymerase chain reaction (PCR) with random

* ozoline@rambler.ru

primers followed by electrophoretic fractionation is used [5, 6]. DNA fragments between primers located at a distance of 100–3000 bp are amplified giving a more complex set of bands compared to PFGE, but electrophoretic fractionation of shorter fragments is much easier compared to long fragments in the previous technique. In the AFLP method (Amplified Fragment Length Polymorphism), the DNA is treated with rare-cutting endonuclease (as in PFGE) and also with another restriction enzyme. The ends of the fragments formed by the first restriction enzyme are specifically ligated by adapters with a known sequence, and the relatively short fragments between them and the restriction sites of the second restriction enzyme are amplified and fractionated [7]. All these methods are widely applied for the typing of pathogenic bacteria, but for their implementation, the genomic DNA of purified bacterial cultures are required, which makes it impossible to use them for metagenomic analysis.

There are several methods that are mainly used to typify individual genomes, but can also be applied to detect particular microorganisms in complex bacterial populations. Thus, for example, the AFLP method in a format when the genome target areas are first amplified with a gene or genome-specific primers, and then subjected to restriction and fractionation [8]. This procedure of experimental processing does not allow obtaining a full-genomic representation of analyzed polymorphisms, but it increases the accuracy of analysis for target sequences and facilitates the interpretation of the obtained patterns.

A particularly large set of methods was designed for targeted analysis of the variable regions of the genome. These can be genes that are present only in the specific strains, antibiotic resistance genes, virulence genes, or housekeeping genes if it is possible to select species-specific primers for them [3, 9–11]. Often, this analysis requires several or even many genomic loci (Multi-Locus Analysis). A special group consists of methods for typing bacteria on the basis of repetitive sequences. It includes the MLVA method (Multi-Locus Variable number tandem repeat Analysis), in which the number of tandem repeats is informative. Experimentally, it can be estimated either by the size of the amplicons [12, 13], or, more precisely, by direct sequencing [14]. Repetitive extragenic palindromes [15], mobile elements [16], CRISPR-Cas cassettes [17], and other genomic features are also used. It is important, however, that for the genomes of each species, the optimal method for typing over variable regions must be selected individually.

The gold standard for the analysis of composite bacterial populations is the typing method for 16S rRNA genes. The results obtained by this approach in different laboratories are stored in public databases [18, 19] and are widely used for phylogenetic analysis. The method is based on the high conservatism of the 16S rRNA genes that are present in all bacterial genomes. The constant regions of these genes are used to select "universal" primers, and 5–9 variable regions as species-specific markers. Applying high-throughput parallel sequencing of 5–9 amplicons obtained from a joint DNA of a complex bacterial community each with an average of 5 copies of the 16S rRNA genes, it is possible to get a library of reads with a very high coverage of the analyzed polymorphic regions, even by using not very powerful sequencers. This makes it possible to receive information not only about the species composition in microbiomes, but also about the number of bacteria of different species in the population. However, this method also has some limitations. The main is the presence in all bacterial genomes of several copies of genes encoding 16S rRNA that may differ in their primary structure [20]. Sometimes intra-genomic variations even exceed interspecies polymorphism. Therefore, the typing according to 16S rRNA sequences allows reliable identification of bacterial genus in more than 90 % of cases, although species identity is established only in 65–83 % of cases [21].

Here we show the results of a pilot project aimed at developing a new genotyping method that has the largest discriminatory power at the species identification level. Our approach uses a large set of unique sequences in genomes as their markers. It is based on total sequencing of

bacterial chromosomes or metagenomes and is ideally suited for bacterial systematization by assessing the phylogenetic proximity of the new isolate to known species by the number of their unique sequences in its genome. The possibility of using a new approach to characterize species diversity in bacterial communities and to assess their relative viability is discussed.

METHODS

Reference database and genomes

A local reference database containing 2443 complete bacterial genomes was created for the study. It was composed from the nucleotide sequences of prokaryotic genomes taken from RefSeq database (ftp://ftp.ncbi.nih.gov/genomes/archive/old_refseq/Bacteria/all.fna.tar.gz) [22] on June 2, 2015 after all chromosomes containing degenerate nucleotides (S, W, R, Y, K, M, B, D, H, V, N) were removed from the initial list. Table 1 provides the information for six genomes of this database used for comparative sequence analysis.

Table 1. List of analyzed genomes

Species name	Accession number in GenBank	Genome length, bp	GC-content, %	Reference
Enterobacter cloacae subsp. *dissolvens* SDM	NC_018079.1	4968248	55.1	[23]
Enterobacter cloacae EcWSU1	NC_016514.1	4734438	54.6	[24]
Enterobacter cloacae subsp. *cloacae* str. ENHKU01	NC_018405.1	4726582	55.1	[25]
Enterobacter lignolyticus SCF1	NC_014618.1	4814049	57	[26]
Clostridium sporogenes DSM 795	CP011663.1	4142990	28	[27]
Cellulomonas flavigena DSM 20109	NC_014151.1	4123179	74.3	[28]

Search for unique sequences in genomes and the strategy of their analysis

The search for unique sequences with 14–28 base pairs (bp) in length was carried out for the test genome using the UniSeq program described in the main text. For this analysis, all genomes belonging to microorganisms of the same genus were removed from the reference database. Scanning was done with 1 bp resolution. Sequences of a particular length found in the analyzed genome, but not in the remaining reference genomes, were considered as unique *k*-mers. Their distribution in the genome relative to the coding regions of genes and intergenic spaces was evaluated using the RefSeq genomic maps [22]. To avoid ambiguity, *k*-mers were sorted into three categories. A particular sequence was considered belonging to the coding region of a gene (or intergenic space) if all its nucleotides were in a given locus. Otherwise, the *k*-mer was attributed to a borderline group.

Identifying discriminative sets of unique sequences

The degrees of intersections between sets of unique *k*-mers found in four *Enterobacter* genomes were estimated using the auxiliary option of the program complex UniSeq. When searching for identical *k*-mers in two sets, the program takes into account the possibility of their presence as complementary sequences. All possible combinations of two and three sets were evaluated, and common sets of unique *k*-mers for all of them were obtained. These data were plotted in Venn diagrams using the Internet resource Venn Diagrams with D3.js [29].

RESULTS AND DISCUSSION

The search of unique *k*-mers in bacterial genomes

A number of methods have been developed for pattern search in large DNA database. These include FASTA [30, 31], BLAST [32], PatternHunter [33], MUMMER [34], SSAHA [35], Fast String Matching Algorithms [36], BWA-SW [37] and others.

The Smith–Waterman-based methods such as FASTA mines approximate matches by employing dynamic programming techniques and are computationally very intensive. BLAST and its variants are an improvement over FASTA in that they use certain seeds for basic anchoring, which are then extended to exact or approximate matches. However, apart from being probabilistic in nature, BLAST type algorithms require large amounts of memory and computing time. PatternHunter is a similar seed-based technique but is still inefficient for applications that involve whole genomes or large databases. Recent suffix tree-based methods [38], such as Mummer, have a very low-search time complexity. They represent all suffixes of the text as a plurality of intermingled linked lists. When the knowledge about genomes gets updated frequently, updating the suffix tree in place becomes tedious, as the inter-mingling of linked lists is very sensitive to changes in the text data. Moreover, since in addition to textual information every node in the tree needs to hold tree-related information, such as pointers to its parents and children apart from text-based information, even the best implementation of suffix trees require about 16 bytes per base [34], which scales up to 46 GB of memory for the preprocessed Human Genome. Deterministic Finite Automaton (DFA)-based methods [38, 39] such as BWA-SW combine DFA and dynamic-programming-based alignment methods. The method does not scale well for large sequences, even for the best case of exact matches. And as they use dynamic programming, the memory requirements of the method are huge. Hashing-based methods such as SSAHA and those proposed by Lecroq [36] offered substring matching and *k*-mers hashing method to greatly improve the time complexity.

To summarize, efficient biological pattern-search algorithms must take into account two problems. First, the possibility of random access to the text, without which the time complexity of the algorithm shoots up to an unacceptable $O(L_G)$ [39], where L_G is the length of the DNA sequence *G*, which is of the order of several billion bases. This can be solved by employing mechanisms such as suffix trees and hash tables. Hashing methods are considered because changing data locally is an easy task when information in the corresponding sequence gets updated. The second problem is related to memory constraints.

In general, the problem of finding unique oligonucleotides of size *k* (*k*-mers) in a database is formalized as follows. There are two non-overlapping sets of DNA sequences (genomes) *E* (test set) and *T* (target set), as well as a subset $S \subseteq E$. We use the term "*k*-mer" to denote a contiguous sequence of DNA bases that is *k* bases long. Each sequence of DNA (genome) *G* that is L_G bases long will contain $(L_G - k + 1)$ overlapping *k*-mers. The task is to find all such *k*-mers that are present in every genome of *S*, but are missing in the target set *T*. To find all *k*-mers in a given DNA sequence $G \in S$, that are absent in *T*, we use an algorithm, which is based on the preprocessing of the database. All the sequences of the reference database were converted into hash tables that efficiently link keys to corresponding values called buckets [40]. The key refers to each distinct *k*-mer while bucket refers to the list of locations of that *k*-mer in the genome. The hash tables were sorted in a special way to minimize the time of calculation while processing the specified *k*-mer.

The straight-line approach, in which every *k*-mer of the genome *G* should be compared to all *k*-mers in sequences of set *T* has the complexity C that lies in the range $N \times \Sigma \le C \le L_G \times \Sigma$, where *N* is the number of unique *k*-mers, Σ – the total length of all targets. For $L_G \approx 5 \cdot 10^6$ bp and $\Sigma = 8008249554 - 2443(k - 1)$ it may require $\sim 10^{16}$ comparisons, which is both time and resource consuming. Thus, the time cost for the search of *k*-mers of length in the range from 14 to 28 was reduced by hashing of all prefixes, which

are 12 bases long. The 12 base pairs for the prefix were selected because no unique sequences of this length were found in any of the tested genomes. The checking the uniqueness of *k*-mer with a prefix *X* in the target set comes down to searching only for the continuation of the prefix. Hence, only the tails located after the prefix are to be compared with the test sequence.

From the preliminary analysis, performed for 29 bacterial genomes (their total length amounts to 2.3 % of the length of full database) taken from different taxa and possessing different GC-content (from 22.5 % to 74.2 %), it became clear that on average only 6.8 % of 12-mers have more than one copy in the genome. If most of these copies give a new unique sequence, this can give 5–6 % additional *k*-mers to the set found only with the first occurrence. Thus, ignoring *k*-mers with the prefix *X*, if other *k*-mer with the same prefix in the genome have already been verified, we had another chance to reduce the time of the search, still obtaining ~95 % of unique *k*-mers.

The Paradox relational database is used to collect the sorted in a special way hash tables containing the offsets of all 12-mers from target genome set. The software UniSeq was developed, which consists of database management system interface and computational procedures with window user interface designed for searching and analyzing unique sequences in bacterial genomes. The database is managed by the use of ObjectPAL operators that are implemented as functions in C++. This allowed creating a window interface that does not require knowledge of database query language from the user, which greatly facilitates the analysis.

Bacterial genomes contain a huge number of short unique sequences

With the development of modern genomic methods, the problem of their full use for monitoring natural communities of microorganisms has become extremely topical. In addition to pathogenic bacteria, for which numerous test systems have already been designed, putrefactive microorganisms that are part of the healthy microbiome also pose a great danger to health if dominate there. The bright representatives of such microorganisms are bacteria of the genus *Clostridium*, which determined the choice of the genome of *Clostridium sporogenes* DSM 795 (*Cl. sporogenes*) as one of the model objects for assessing the potential of the new genotyping method. The genome of this bacterium has an anomalously high content of A/T bp (Table 1), which could not but affect the results of the planned work. Therefore, the genomes of three strains of the conditionally pathogenic species of *Enterobacter cloacae* (*E. cloacae*) widely distributed in nature and being a commensal of our intestinal tract, were also used for the analysis. The corresponding genomes have practically the same content of A/T and G/C pairs. The strain of *Enterobacter lignolyticus* (*E. lignolyticus*) initially annotated as *E. cloacae* on the basis of 16S rRNA typing and later classified as an independent species on the basis of multi-locus sequence typing [26] was added as an experimental sample to check discriminative capacity of the software. Finally, the genome of *Cellulomonas flavigena* DSM 20109 (*C. flavigena*) with an abnormally high GC-content was also used for analysis as a compositional antipode for the genome of *Cl. sporogenes*.

Each of the six genomes was scanned by UniSeq using a reference database for comparison and variable size of *k*-mers. As expected, the number of unique sequences increased with increasing length *k* and their total number turned out to be very large (solid lines in Fig. 1,A), which characterizes the degree of uniqueness of genomes. The greatest number of unique sequences was found in the genome of *E. cloacae* SDM, but the sigmoidal curve tracking their dependence on *k* was almost the same as for the two other *Enterobacter* genomes (*E. cloacae* EcWSU1 and *E. cloacae* ENHKU01) with completely overlapping plots (black solid lines). In full accordance with the current classification, the curve for *E. lignolyticus* SCF1 clearly differed from them (red solid curve), thereby assuming the ability of our approach to detect even a subtle interspecies difference that was not evident on

the basis of 16S rRNA typing and was found only by the method of multi locus sequence typing [26, 41].

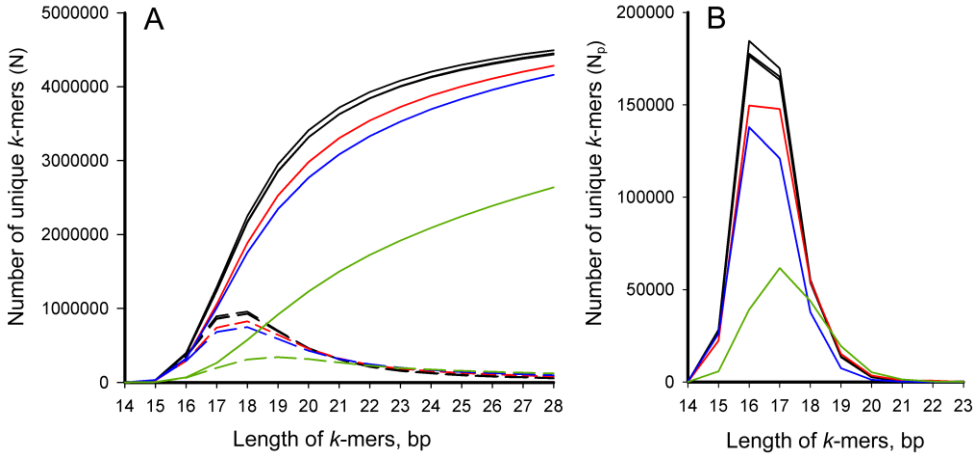

Fig. 1. The dependence of the number of unique *k*-mers on their length in the model genomes. **A**: solid lines – cumulative curves, dashed lines – increment curves plotted for $\Delta N/\Delta k$. **B** – differential plots for "primary" *k*-mer (specified in the text). Black color were used for the genomes of three *E. cloacae* strains (SDM, EcWSU1, ENHKU01), red for *E. lignolyticus* SCF1, blue for *Cl. sporogenes* DSM 795 and green for *C. flavigena* DSM 20109.

In the A/T-rich genome of *Cl. sporogenes* the number of the unique sequences turned out to be smaller (blue line) and much smaller in the G/C-rich genome of *C. flavigena*. The deviation of the corresponding plots from the *E. cloacae* curves representing genomes with ~55 % GC-content was not proportional to the difference in their base composition. The degree of uniqueness of genomes, therefore, is determined not only by the imbalance in their A(T) and G(C) composition.

The largest contribution to the number of unique sequences was given by the 17–19-mers (dashed lines in Fig. 1,A). However, most of them are derived from shorter unique *k*-mers. If, for instance, the 16-mer ATGCCGTTAATTAAAG from the test genome was absent in the database and, accordingly, was rated as "unique", any 17-mer with the same prefix will be also "unique". On the other hand, if the 15-mer ATGCCGTTAATTAAA was found in the database, the indicated 16-mer can be considered as a "primary unique" sequence. Most of such "seed" sequences in all genomes are 16- or 17-mers (Fig. 1,B). Thus, unique *k*-mers in the range of length 16–18 bases can be suggested as potential markers for typing.

Table 2. The distribution of unique *k*-mers relative to the coding sequences

Name of the species	Type of data	Coding density of the genome (%)	Percentage of unique *k*-mers		
			genes	intergenic spaces	borderline *k*-mers
Enterobacter cloacae subsp. *dissolvens* SDM	Np	88.96	81.25	16.80	1.95
	N		78.97	18.81	2.22
Enterobacter cloacae EcWSU1	Np	90.27	83.40	14.80	1.80
	N		81.07	16.81	2.12
Enterobacter cloacae subsp. *cloacae* str. ENHKU01	Np	89.02	81.65	16.48	1.87
	N		79.29	18.53	2.18
Enterobacter lignolyticus SCF1	Np	89.58	78.78	18.97	2.25
	N		75.75	21.50	2.75
Clostridium sporogenes DSM 795	Np	81.83	84.38	14.62	1.00
	N		83.74	15.2	1.06
Cellulomonas flavigena DSM 20109	Np	90.35	79.81	17.61	2.58
	N		67.29	28.59	4.11

Unique *k*-mers are almost proportionally distributed between genes and intergenic spaces

The largest part of the currently used phylotyping protocols relies on the conservatism of the coding sequences of discriminating genes, such as 16S rRNA genes, virulence genes or toxicity, whose intraspecies stability is controlled by evolutionary selection. The other part, on the contrary, for characterizing intraspecies polymorphism, uses highly variable sequences of intergenic regions, such as tandem repeats, REP elements or CRISPR cassettes. Thus, the area of effective application of a new approach depends on how unique *k*-mers are distributed between genes and intergenic spaces. This analysis was performed for all *k*-mers in the range of length from 14 to 28 bases for all *Enterobacter* strains and *C. flavigena*, and the bias towards the intergenic spaces was observed. We did not detect such bias for *Cl. sporogenes* that has anomalously low coding density. In Table 2, the data for primary (Np) and cumulative (N) sets of 16-mers are given as examples. Thus, it became clear that less conservative intergenic spaces may be the places where genomic signatures are predominantly generated.

UniSeq-based search for genomic signatures revealed strain- and species-specific sets for discriminatory phylotyping

To assess the evolutionary stability and plasticity of bacterial genomes, the degrees of overlap between the obtained sets of primary and cumulative *k*-mers unique for each of the three genomes of *E. cloacae* and the genome of *E. lignolyticus*, were evaluated. Figure 2 exemplifies the data obtained for 16-mers. As expected, the common sets for genomes of the same species were much larger (Fig. 2,A) than the common sets of identical *k*-mers in the genomes of *E. cloacae* and *E. lignolyticus* (Fig. 2,B). Surprisingly, each set turned out to be mostly composed of the strain-specific sequences. The part of common *k*-mers in the genomes of two strains belonging to *E. cloacae* species was only 9–14 % and increased with decreasing *k*. However, the common sets of unique *k*-mers for the three strains decreased by only ~50 % compared to the common sets of two strains and included more than seventeen thousand potential markers. If this tendency to preserve the core part persists with an increase in the number of sequenced genomes of *E. cloacae*, it can be directly used for typing new isolates. Otherwise, their species identity can be judged by the presence in the genome of unique *k*-mers found in other strains of the same species.

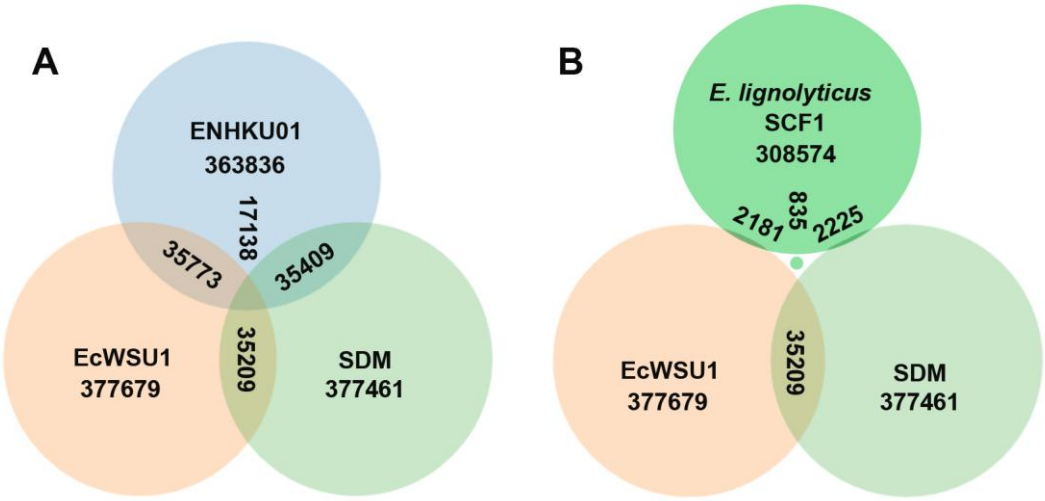

Fig. 2. Venn diagrams for intersection of cumulative sets of unique 16-mers. Panel **A** shows diagram obtained for genomes of three *Enterobacter* species. Panel **B** exemplifies a lesser intersection of the sets found in the genome of *E. lignolyticus* and genomes of two species of *Enterobacter cloacae* with a lager mutual deviation. Numerals indicate the size of the sets, as well as their common parts.

The same strategy can be offered to identify the genus of novel isolate, since the number of common *k*-mers found in the genomes of *E. cloacae* and *E. lignolyticus* is more than an order of magnitude smaller (Fig. 2,B). The small genus-specific sets also show a certain tendency to preserve the core part: the number of unique *k*-mers common to genomes of *E. cloacae* strains and *E. lignolyticus* (2181/2225) compose only 0.5–0.7 % of their quantity in each genome, whereas the percentage of common *k*-mers in three genomes (835) is only 3-fold smaller. It is clear that with the addition of new species, the size of "genus-specific" set will further decrease, making the use of such *k*-mers unreliable. The lesser discriminative capacity for genus was not *a priori* expected, but may reflect the different impact given by intraspecies stability controlled by evolutionary selection and intrageneric divergence required for speciation.

CONCLUSION

A special feature of the post-genomic era is the great need for precise phylogenetic typing of bacteria as biological objects and in methods that allow characterizing the diversity of bacterial communities. As a result, the methods of multi locus typing are widely used to overcome the limitations of the classical method of identifying microorganisms based on 16S rRNA polymorphism. It is already obvious that with the use of direct sequencing these methods provide much more information than in the PCR format [42] and network resource has appeared that facilitates the systematics of new isolates using short sequences (400–500 bp) of multiple housekeeping genes [43]. In this study, we report the first results obtained by the typing method based on UniSeq program complex that operates only with *k*-mers, without any link to a particular group of genes and potentially can work with incomplete genomes, contigs or even shotgun sequences. It is likely that the discriminative capacity of the approach makes it possible to distinguish the genomes of different strains even better than genomes of different species. This approach, therefore, can be used as a complementary technique to classical genotyping according to 16S rRNA genes that more reliably identifies the genus of new isolates than its species.

Although special efforts are required to assess the minimum number of unique *k*-mers needed to determine a species, it is important that they are all equivalent in their "marker capability", and any combination is informative. In this study, we showed that the method perfectly discriminates bacteria from different species. This is of particular importance for pathogenic and toxic microorganisms. Genotyping of bacterial communities is much more difficult. If, for example, it is necessary to characterize microbiome consisting of 1000 different bacterial species, then using medium-class sequencers such as Illumina MiSeq or Ion Torrent PGM, it is possible to get 5–6000000 high-quality sequence reads with a length of ~150 bases. This yields on average ~800000 16-mers, per genome, covering approximately 20 % of its sequence. Depending on base composition, ~10000–64000 of these *k*-mers will be unique (Fig. 1,A) and ~900–6000 such sequences will overlap with a set of unique *k*-mers in the genomes of known species of the same genus (Fig. 2,A). This is much more than gives any other multi locus typing, even if difference in the presence of particular species varies over a wide range. The loss in coverage is compensated by the high multiplicity of targeted regions and 10–20 % of losses due to sequencing errors do not seem dramatic.

Working with very short sequences, the method does not require the assembly of target loci and can be used to analyze meta-transcriptomes, thereby providing information about living microorganisms and their functionality. However, there are at least two shortcomings that complicate the implementation of the new technique in a widely available resource. First, for comparison, the most complete and promptly updated reference database is required. Secondly, to search for unique *k*-mers, it is necessary to remove from this base all genomes of the same genus. Thus, preliminary studies are required to identity the genus. Alternatively, a strategy for the sequential seizure of genomes belonging to ~ 2500 genera from a common

database can be projected, but it is very time-consuming. Creation a parallel database containing sets of unique *k*-mers for different species can certainly facilitate the implementation of the method.

The need for direct sequencing is no longer considered as a large limitation. As the cost of high-throughput sequencing is rapidly declining, it is becoming increasingly available for routine laboratory practice, especially since new inexpensive and highly processive sequencers from Oxford Nanopore Technologies will soon be available. It is important that the data obtained by direct sequencing do not depend on experimental protocols, the quality of gel fractionation or the markers used. They are self-contained and termless, but the use of the valuable genomic data accumulated in public resources for mixed bacterial populations requires new approaches. One such approach is proposed in this study.

The study was partially supported by the Russian Foundation for Basic Research (grants №16-04-01570 and №15-07-05783).

REFERENCES

1. Sabat A.J., Budimir A., Nashev D., Sa-Leao R., van Dijl J.M., Laurent F., Grundmann H., Friedrich A.W., on behalf of the ESCMID Study Group of Epidemiological Markers (ESGEM). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveill*. 2013. V. 18. Article No. 20380. doi: 10.2807/ese.18.04.20380-en.
2. Skolotneva E.S., Volkova R.A., Elbert E.V., Mironov A.N., Merkulov V.A., Bondarev V.P., Borisevich I.V. Bacterial genotyping methods: banding pattern-based analysis. *Biopreparation* (*Biopharmaceuticals*). 2014. V. 2. P. 13–21.
3. Bondareva O.S., Savchenko S.S., Tkachenko G.A., Abueva A.I., Muratova Yu. O., Antonov V.A. Modern approaches to genotyping of causative agents of particularly dangerous infections. *Epidemiology and infectious diseases*. 2014. V. 1. P. 35–43.
4. Swaminathan B., Barrett T.J., Hunter S.B., Tauxe R.V. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg. Infect. Dis*. 2001. V. 7. P. 382–389. doi: 10.3201/eid0703.010303
5. Lanini S., D'Arezzo S., Puro V., Martini L., Imperi F., Piselli P. Montanaro M., Paoletti S., Visca P., Ippolito G. Molecular epidemiology of a *Pseudomonas aeruginosa* hospital outbreak driven by a contaminated disinfectant-soap dispenser. *PloS One*. 2011. V. 6. Article No. e17064. doi: 10.1371/journal.pone.0017064
6. Chang H.L., Tang C.H., Hsu Y.M., Wan L., Chang Y.F., Lin C.T., Tseng Y.R., Lin Y.J., Sheu J.J., Lin C.W., Chang Y.C., Ho M.W., Lin C.D., Ho C.M., Lai C.H. Nosocomial outbreak of infection with multidrug-resistant *Acinetobacter baumannii* in a medical center in Taiwan. *Infect. Control Hosp. Epidemiol*. 2009. V. 30. P. 34–38. doi: 10.1086/592704
7. Zhao S., Mitchell S.E., Meng J., Kresovich S., Doyle M.P., Dean R.E., Casa A.M., Weller J.W. Genomic typing of *Escherichia coli* O157:H7 by semi-automated fluorescent AFLP analysis. *Microbes Infect.* 2000. V. 2. P. 107–113. doi: 10.1016/S1286-4579(00)00278-1
8. Chowdhury N., Asakura M., Neogi S.B., Hinenoya A., Haldar S., Ramamurthy T., Sarkar B.L., Faruque S.M., Yamasaki S. Development of simple and rapid PCR-fingerprinting methods for *Vibrio cholerae* on the basis of genetic diversity of the superintegron. *J. Appl. Microbiol*. 2010. V. 109. P. 304–312. doi: 10.1111/j.1365-2672.2009.04658.x
9. Li Y., Dai E., Cui Y., Li M., Zhang Y., Wu M., Zhou D., Guo Z., Dai X., Cui B., Qi Z., Wang Z., Wang H., Dong X., Song Z., Zhai J., Song Y., Yang R. Different region

analysis for genotyping *Yersinia pestis* isolates from China. *PLoS One*. 2008. V. 3. Article No. e2166. doi: 10.1371/journal.pone.0002166

10. Duangsonk K., Gal D., Mayo M., Hart C.A., Currie B.J., Winstanley C. Use of a variable amplicon typing scheme reveals considerable variation in the accessory genomes of isolates of *Burkholderia pseudomallei*. *J. Clin. Microbiol*. 2006. V. 44. P. 1323–1334. doi: 10.1128/JCM.44.4.1323-1334.2006
11. Huber B., Scholz H.C., Lucero N., Busse H.J. Development of a PCR assay for typing and subtyping of *Brucella* species. *Int. J. Med. Microbiol.* 2009. V. 299. P. 563–573. doi: 10.1016/j.ijmm.2009.05.002
12. Sabat A., Krzyszton-Russjan J., Strzalka W., Filipek R., Kosowska K., Hryniewicz W., Travis J., Potempa J. New method for typing *Staphylococcus aureus* strains: multiple-locus variable-number tandem repeat analysis of polymorphism and genetic relationships of clinical isolates. *J. Clin. Microbiol*. 2003. V. 41. P. 1801–1804. doi: 10.1128/JCM.41.4.1801-1804.2003
13. Elberse K.E., Nunes S., Sa-Leao R., van der Heide H.G., Schouls L.M. Multiple-locus variable number tandem repeat analysis for *Streptococcus pneumoniae*: comparison with PFGE and MLST. *PloS One.* 2011. V. 6. Article No. e19668. doi: 10.1371/journal.pone.0019668
14. Visca P., D'Arezzo S., Ramisse F., Gelfand Y., Benson G., Vergnaud G., Fry N.K., Pourcel C. Investigation of the population structure of *Legionella pneumophila* by analysis of tandem repeat copy number and internal sequence variation. *Microbiology*. 2011. V. 157. P. 2582–2594. doi: 10.1099/mic.0.047258-0
15. Wilson M.K., Lane A.B., Law B.F., Miller W.G., Joens L.A., Konkel M.E., White B.A. Analysis of the pan genome of *Campylobacter jejuni* isolates recovered from poultry by pulsed-field gel electrophoresis, multilocus sequence typing (MLST), and repetitive sequence polymerase chain reaction (rep-PCR) reveals different discriminatory capabilities. *Microb. Ecol*. 2009. V. 58. P. 843–855. doi: 10.1007/s00248-009-9571-3
16. McCarthy A.J., Breathnach A.S., Lindsay J.A. Detection of mobile-genetic-element variation between colonizing and infecting hospital-associated methicillin-resistant *Staphylococcus aureus* isolates. *J. Clin. Microbiol*. 2012. V. 50. P. 1073–1075. doi: 10.1128/JCM.05938-11
17. Liu F., Kariyawasam S., Jayarao B.M., Barrangou R., Gerner-Smidt P., Ribot E.M., Knabel S.J., Dudley E.G. Subtyping *Salmonella enterica* serovar *enteritidis* isolates from different sources by using sequence typing based on virulence genes and clustered regularly interspaced short palindromic repeats (CRISPRs). *Appl. Environ. Microbiol*. 2011. V. 77. P. 4520–4526. doi: 10.1128/AEM.00468-11
18. Wang Q., Garrity G.M., Tiedje J.M., Cole J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 2007. V. 73. P. 5261–5267. doi: 10.1128/AEM.00062-07
19. DeSantis T. Z., Hugenholtz P., Larsen N., Rojas M., Brodie E. L., Keller K., Huber T., Dalevi D., Hu P., Andersen G. L. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol*. 2006. V. 72. P. 5069–5072. doi: 10.1128/AEM.03006-05
20. Větrovsky T., Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One*. 2013. V. 8. Article No. e57923. doi: 10.1371/journal.pone.0057923
21. Andersson A.F., Lindberg M., Jakobsson H., Backhed F., Nyren P., Engstrand L. Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One*. 2008. V. 3. Article No. e2836. doi: 10.1371/journal.pone.0002836

22. Tatusova T., Ciufo S., Fedorov B., O'Neill K., Tolstoy I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* 2014. V. 42. P. D553–D559. doi: 10.1093/nar/gkt1274
23. Xu Y., Wang A., Tao F., Su F., Tang H., Ma C., Xu P. Genome sequence of *Enterobacter cloacae* subsp. *dissolvens* SDM, an efficient biomass-utilizing producer of platform chemical 2,3-butanediol. *J. Bacteriol.* 2012. V. 194. P. 897–898. doi: 10.1128/JB.06495-11
24. Humann J.L., Wildung M., Cheng C.-H., Lee T., Stewart J.E., Drew J.C., Triplett E.W., Main D., Schroeder B.K. Complete genome of the onion pathogen *Enterobacter cloacae* EcWSU1. *Stand. Genomic Sci.* 2011. V. 5. P. 279–286. doi: 10.4056/sigs.2174950
25. Liu W.-Y., Chung K. M.-K., Wong C.-F., Jiang J.-W., Hui R. K.-H., Leung F. C.-C. Complete genome sequence of the endophytic *Enterobacter cloacae* subsp. *cloacae* strain ENHKU01. *J. Bacteriol.* 2012. V. 194. P. 5965. doi: 10.1128/JB.01394-12
26. DeAngelis K.M., D'Haeseleer P., Chivian D., Fortney J.L., Khudyakov J., Simmons B., Woo H., Arkin A.P., Davenport K.W., Goodwin L., Chen A., Ivanova N., Kyrpides N.C., Mavromatis K., Woyke T., Hazen T.C. Complete genome sequence of "*Enterobacter lignolyticus*" SCF1. *Stand. Genomic Sci.* 2011. V. 5. P. 69–85. doi: 10.4056/sigs.2104875
27. Nakano K., Terabayashi Y., Shiroma A., Shimoji M., Tamotsu H., Ashimine N., Ohki S., Shinzato M., Teruya K., Satou K., Hirano T. First complete genome sequence of *Clostridium sporogenes* DSM 795^{T}, a nontoxigenic surrogate for *Clostridium botulinum*, determined using PacBio single-molecule real-time technology. *Genome Announc.* 2015. V. 3. Article No. e00832-15. doi: 10.1128/genomeA.00832-15
28. Abt B., Foster B., Lapidus A., Clum A., Sun H., Pukall R., Lucas S., Glavina Del Rio T., Nolan M., Tice H., Cheng J.F., Pitluck S., Liolios K., Ivanova N., Mavromatis K., Ovchinnikova G., Pati A., Goodwin L., Chen A., Palaniappan K., Land M., Hauser L., Chang Y.J., Jeffries C.D., Rohde M., Goker M., Woyke T., Bristow J., Eisen J.A., Markowitz V., Hugenholtz P., Kyrpides N.C., Klenk H.P. Complete genome sequence of *Cellulomonas flavigena* type strain (134). *Stand. Genomic Sci.* 2010. V. 3. P. 15–25. doi: 10.4056/sigs.1012662
29. Frederickson B. *Venn Diagrams with D3.js*. URL: http://www.benfrederickson.com/venn-diagrams-with-d3.js (accessed 10.11.2017).
30. Lipman D.J., Pearson W.R. Rapid and sensitive protein similarity searches. *Science*. 1985. V. 227. P. 1435–1441. doi: 10.1126/science.2983426
31. Pearson W.R., Lipman D.J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* 1988. V. 85. P. 2444–2448. doi: 10.1073/pnas.85.8.2444
32. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J. Mol. Biol.* 1990. V. 215. P. 403–410. doi: 10.1016/S0022-2836(05)80360-2
33. Ma B., Tromp J., Li M. PatternHunter: faster and more sensitive homology search. *Bioinformatics*. 2002. V. 18. P. 440–445. doi: 10.1093/bioinformatics/18.3.440
34. Kurtz S., Phillippy A., Delcher A.L., Smoot M., Shumway M., Antonescu C., Salzberg S.L. Versatile and open software for comparing large genomes. *Genome Biol.* 2004. V. 5. Article No. R12. doi: 10.1186/gb-2004-5-2-r12
35. Ning Z., Cox A.J., Mullikin J.C. SSAHA: a fast search method for large DNA databases. *Genome Res.* 2001. V. 11. P. 1725–1729. doi: 10.1101/gr.194201
36. Lecroq T. Fast exact string matching algorithms. *Inf. Process. Lett.* 2007. V. 102. P. 229–235. doi: 10.1016/j.ipl.2007.01.002
37. Li H., Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2010. V. 26. P. 589–595. doi: 10.1093/bioinformatics/btp698

38. Gusfield D. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge: Cambridge University Press, 1997.
39. Charras C., Lecroq T. *Handbook of Exact String Matching Algorithms*. London: Kings College, 2004.
40. Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. In: *Introduction to Algorithms*. Cambridge, London: MIT Press, 2009. P. 253–286.
41. Liu W.-Y., Wong C.-F., Chung K.-M., Jiang J.-W., Leung F.C.-C. Comparative genome analysis of *Enterobacter cloacae*. *PLoS One.* 2013. Article No. e74487. doi: 10.1371/journal.pone.0074487
42. Larsen M.V., Cosentino S., Rasmussen S., Friis C., Hasman H., Marvig R.L., Jelsbak L., Sicheritz-Pontén T., Ussery D.W., Aarestrup F.M., Lund O. Multilocus sequence typing of total genome sequenced bacteria. *J. Clin. Micobiol.* 2012. V. 50. P. 1355–1361. doi: 10.1128/JCM.06094-11
43. MLST 1.8 (MultiLocus Sequence Typing): Home Page. URL: https://cge.cbs.dtu.dk/services/MLST/ (accessed 10.11.2017).

Received November 25, 2017.
Published December 19, 2017.