 doi:10.20944/preprints202201.0296.v1

Article

CRISPR-Cas Systems in Gut Microbiome of children with Autism Spectrum Disorders

N.V. Zakharevich [1,†], M.S. Nikitin [1,2,†], A.S. Kovtun [1,3], V.O. Malov [4], O.V. Averina [1], V.N. Danilenko [1,2], I.I. Artamonova [1,5,*]

[1] Vavilov Institute of General Genetics Russian Academy of Sciences, Moscow, Russia
[2] Moscow Institute of Physics and Technology, State University, Dolgoprudny, Russia
[3] Skolkovo Institute of Science and Technology, Skolkovo, Russia
[4] Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia
[5] Kharkevich Institute for Information Transmission Problems, RAS, Moscow, Russia
* Correspondence: irenart@gmail.com
† These authors contributed equally to this work

Abstract: Human gut microbiome is associated with various diseases, including autism spectrum disorders (ASD). Variations of the taxonomical composition in the gut microbiome of children with ASD have been observed repeatedly. However, features and parameters of the CRISPR-Cas systems in the gut microbiome of children with ASD have not been investigated yet. Here we demonstrate such an analysis in comparison with the healthy microbiome. For the identification of CRISPR-Cas systems, we used a combination of the publicly available tools suited for completed genomes with subsequent filtrations. In all considered datasets, the microbiomes of children with ASD contained fewer arrays per Gb of assembly, than the control group, but the arrays included more spacers on average. These patterns were observed systematically in our datasets, although their statistical significance hardly matched the thresholds. CRISPR arrays from the microbiomes of children with ASD differed from the control group neither in the fractions of spacers with protospacers from known genomes, nor in the sets of known bacteriophages providing protospacers. The majority of bacterial protospacers of the gut microbiome systems for both children with ASD and the healthy ones was located in the prophage islands.

Keywords: microbiome; autism spectrum disorders; CRISPR-Cas; protospacer

1. Introduction

Among all possible natural communities suitable for metagenomics studies, human microbiome deserves special attention due to its medical importance. It is becoming clear that microbes populating human body are involved in different processes important for the host and closely linked to his health. Among them are so sophisticated ones as formation of the host's immune system [1] and regulation of metabolic processes [2]. Gut microbiota also can affect neurological functions and even the behavior of the host [3]. Thus, the emergence of more and more evidence for the microbiome association with various diseases comes as no surprise.

For quite a long time, human microbiome has been unattainable for complex experimental studies as a considerable fraction of bacterial species is uncultivated. According to the recent estimations, in metagenomes this fraction makes up as much as 72% of bacterial and 69% of archaeal species [4]. Only forty years ago new methods allowed sequencing of the genes encoding 5S [5] and 16S RNAs [6], which became a new milestone in taxonomic analysis of the microbiome [7]. For the whole genomes of organisms shaping the human microbiome, the problem was solved in principle only with the introduction of the next generation sequencing technologies, when the direct sequencing of natural environmental samples had become possible. Massive sequencing of human microbiome

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 January 2022 doi:10.20944/preprints202201.0296.v1

was targeted by several international projects, such as Human Microbiome Project [8] and MetaHIT [9], and a series of national ones [10-12]. It was shown that the taxonomic composition of the sequenced samples varied dramatically among different parts of the human body they originated from and was changing gradually among them. Due to that and based on the simplicity of probe obtaining, it became customary to discuss microbiomes of separate human organs or cavities. For instance, within the HMP projects several probes from each donor were sequenced, – namely nasal, oral, vaginal, gut and skin microbiomes. Gut (or intestinal) microbiome turned out to be highly diverse and became the most popular one for investigations.

Development of the next generation sequencing technologies and the metagenomic approach made it possible not only to describe the taxonomic composition of the studied microbiomes, but also to evaluate the impact of pathogens on human health, and to study complex changes in the entire microbial community associated with specific diseases. For example, for the gut microbiome, a complex relationship between taxonomic distribution and a variety of diseases was demonstrated. Among them were Crohn disease [13], Ulcerative colitis [14], obesity [15] and others. In addition, further research revealed a complex relationship between the gut microbiome and the brain. Now this two-way relationship is usually called the microbiome-gut-brain axis, and it is being actively studied [16]. Recent research in this area shows a link between neurological diseases such as Alzheimer's [17] and Parkinson's [18] diseases, and autism spectrum disorder (ASD).

The first evidence for the relationship between the gut microbiome and ASD appeared at the turn of the millennium, when the development of regressive ASD as a response to the antibiotic treatment for chronic otitis had been documented [19]. It was suggested that Clostridiales bacteria might be associated with ASD due to the neurotoxin production [20]. Later this suggestion was supported by several studies [21,22], including one that demonstrated the behavioral improvements for ASD children treated with Clostridiales-targeting antibiotics [23]. Finally, it was shown that stool samples from children with ASD contained different types of Clostridiales compared to neurotypical patients [24], with *Clostridium (Lachno-clostridium) bolteae* being typical for children with ASD and gastrointestinal disorders [25]. In addition to *Clostridium*, various works also described such marker genera of autism as *Nitriliruptor*, *Youngiibacter*, *Burkholderia*, *Bilophila*, *Constrictibacter*, *Dichelobacter*, *Bacteroides* and *Prevotella*, as well as two orders – Desulfovibrionales and Methanomicrobiales [26,27].

However, some observations of the changes in gut microbiota of patients with ASD may look contradictory. For example, there is an evidence that composition of ASD patients' microbiome differed significantly in comparison with the healthy ones, showing lower abundances of *Bifidobacterium* species and higher abundances of *Lactobacillus* species [28], or lower abundances of the genera *Prevotella*, *Coprococcus*, and unclassified Veillonellaceae in autistic samples [29]. Additionally, several other studies demonstrated reduced overall richness of microbiome in ASD patients compared to the neurotypical group [30,31]. On the other hand, there were studies, where no significant difference [32], or even higher richness of ASD patients gut microbiome had been observed [33]. But nothing was reported about the role of bacteriophages in the process of the microbiome changes in the course of ASD.

CRISPR-Cas systems (Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated proteins) are coded in special loci of bacteria and archaea. CRISPR arrays combine tandemly repeated short fragments interspersed with commensurate sequences unique for the genome. The latter are called 'spacers' and often demonstrate similarity to genetic mobile elements like phages or plasmids. This feature reflects the spacers' origin and provides the systems with their main function to defend against invading nucleic acids [34]. The unit of a new spacer and an additional copy of the direct repeat is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 January 2022 doi:10.20944/preprints202201.0296.v1

inserted between the leader sequence (a genome segment separating the array from the *cas* genes) and the array, if the host cell has been able to repel an attack of an invader. Thus, the array represents a unique record of the last infections of the particular cell or its vertical descendants in the chronological order [35].

Whereas cultivated bacteria served as a main object for investigations of CRISPR-Cas systems, metagenomics data had recently begun to be in demand in this area. Several studies were conducted from the description of systems in different ecological niches (e.g., [36]) to targeted search for new systems with editing potential (e.g., [37]). Human microbiome was not set aside here as well, being studied with different approaches for arrays' identification and analysis (e.g., [38,39]). In all these studies an emphasis was made on the natural CRISPR-Cas systems in healthy human microbiomes.

In this study, we tried to compare the CRISPR-Cas systems in the gut microbiomes of children with ASD and the control group in order to trace disease-mediated changes in their arrays, if any. The deeper analysis of protospacers became the second goal of this study, because it could not be ruled out that CRISPR arrays contain preferably traces of encounters with specific groups of viruses during ASD. Finally, as almost all previous research had focused on the inhabitants of a healthy microbiota, the possibility of discovering previously unknown CRISPR-Cas systems in a disease-modified human microbiome also seemed promising.

2. Data and Algorithms

The raw data used for the study were downloaded from the NCBI BioProject collection (PRJNA516054). These data were initially obtained by sequencing of faecal probes from 77 children aged 1 to 9 years old, 54 of which were diagnosed with autism spectrum disorders according to DSM-V criteria (Diagnostic and Statistical Manual of mental disorders, fifth edition) [40]. In the original study sequencing was carried out in three independent series of experiments with different protocols on different platforms, that is illustrated in Table 1. No other differences in the parameters of the sample preparation and sequencing among series were indicated in the original manuscript [40]. The downloaded data were processed and assembled as described in [40]. Contigs with length of less than 200 nt were dropped.

Table 1. Description of samples and statistics of sequencing and assembly, important for this study (see [40]).

	Human gut metagenome samples (NCBI BioProject ID: PRJNA516054)		
	Series I	**Series II**	**Series III**
Gender	both sexes		
Age (y.o.)	1 - 9 (ASD) 3 - 4 (control)	2 - 4 (ASD) 2 - 4 (control)	2 - 6 (ASD) 3 (control)
Number of samples	14 (ASD) 5 (control)	15 (ASD) 15 (control)	25 (ASD) 3 (control)
Platform and type of sequencing	Illumina HiSeq 2500, paired-end	Illumina HiSeq 4000, paired-end	Illumina NovaSeq 6000, paired-end
Read length, nt	135	150	150
Range of assembly size, Gb	0.06 – 0.26 (ASD) 0.13 – 0.19 (control)	0.11 – 0.30 (ASD) 0.14 – 0.28 (control)	0.11 – 0.37 (ASD) 0.17 – 0.27 (control)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 January 2022 doi:10.20944/preprints202201.0296.v1

For identification of CRISPR arrays we started from the procedure described in [36] and modified it by changing CRISPRFinder to CRISPRCasFinder [41]. Its aim was to reduce false positive results produced by any of the tools used here. For that, only arrays predicted with all algorithms simultaneously or supported by one of the known biological properties of the real arrays, like co-localization with *cas* genes or similarity of the repeat sequence with other direct repeats, were selected for the analysis. For clustering of direct repeats for the arrays found by any of three tools we used the DNACLUST software [42] with similarity threshold of 0.8. The resulted algorithm is illustrated in Figure 1. All procedures were performed for different series separately.

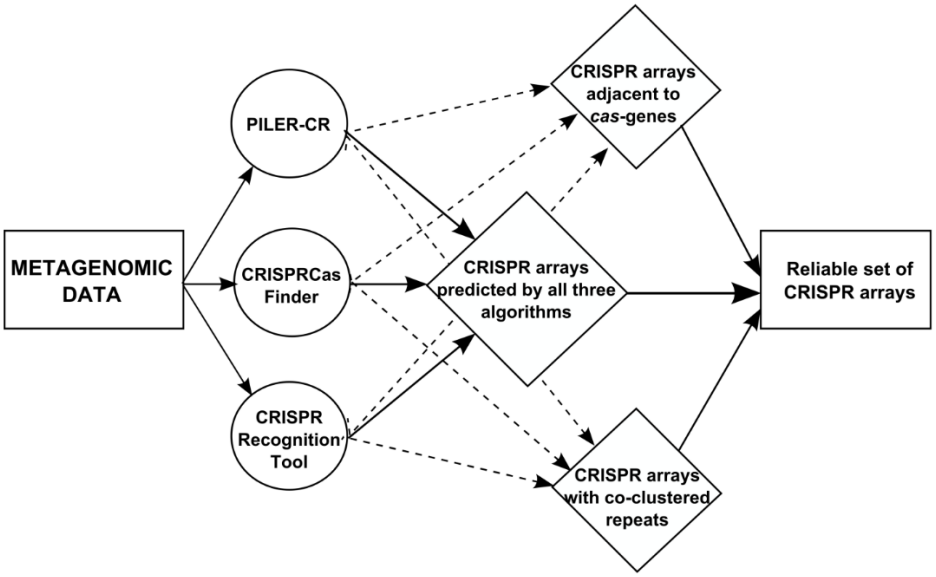


Figure 1. Schematic illustration for the identification of reliable CRISPR arrays (see[36]).

In order to test whether differences in parameters of CRISPR-Cas systems and their distribution were significant between microbiomes of children with ASD and the control group, we compared the observation values for the parameters between datasets of ASD and the control for each series. For that we used the Shapiro-Wilk test for normality and Welch's *t*-test for equality of expectations implemented in the Python scipy.stats package. For single comparisons the significance level of 0.05 was selected. As the test for normality may provide unreliable results for small samples, we also performed the Mann-Whitney-Wilcoxon nonparametric test to confirm the results obtained with the Welch's *t*-test. It was performed with the same significance level using its implementation in the Python scipy.stats package to confirm the results obtained with the Welch's *t*-test.

The arrays were treated as complete, if they were flanked by at least 200 nucleotides on both sides in their contigs. To test the possibility of combining the data of the datasets, we compared the respective values for ASD or the control for different series using the Welch's *t*-test and confirmed the conclusions with the Mann-Whitney-Wilcoxon test.

To search for the disease markers among direct repeats, we selected clusters containing repeats from individual samples of children with ASD and not from the control group. Only precise identity of sequences was allowed in the similar procedure for spacers.

The CRISPR-Cas systems of *Enterocloster bolteae* were identified using CRISPRCasdb [43] and its utilities in two known strains of the species. To investigate whether these systems might be used as the disease markers, we searched for the respective repeats in the microbiome data using the BLAST package [44].

To check whether microbiome strains of *Enterocloster bolteae* could contain CRISPR-Cas systems with other repeats, we annotated all metagenomic contigs with MMseqs2 [45] and selected those assigned to the species. The results of the general procedure of arrays

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 January 2022 doi:10.20944/preprints202201.0296.v1

identification for these contigs were analysed in addition to independent results of CRISPRCasFinder [41].

To find protospacers among bacteria and phages, we aligned spacer sequences to the bacterial and viral sections of the RefSeq database using BLAST [46]. To exclude hits to CRISPR arrays in bacteria, we also aligned the sequences of the respective direct repeats with the same procedure. If a spacer and the corresponding repeat had been aligned to the same bacterium, such case was excluded from the consideration. For each remained hit the alignment was extended up to the full coverage of the spacer using the in-house python script. After that, for the final set of protospacers we selected hits with not more than four mismatches only.

We also analysed the locations of the protospacers in known bacterial genomes. For that purpose, we considered assemblies of bacterial genomes from the genomic subsection of the RefSeq database. If an assembly contained protospacers for arrays from more than ten metagenomic samples from any of the datasets, it was selected for further consideration. For the selected assemblies we identified regions of potentially phage origin with two tools – ProphageHunter [47] and Phast [48], with default parameters for both. If location of a protospacer had not been included into such a region, we analysed the neighbouring gene(s) using their annotation and/or best blast hits of their products. The location was marked as «prophage», if it had been included in a prophage region by at least one of the tools, or the annotation of the respective protein(s), or any of its (their) best blast hits indicated the phage origin.

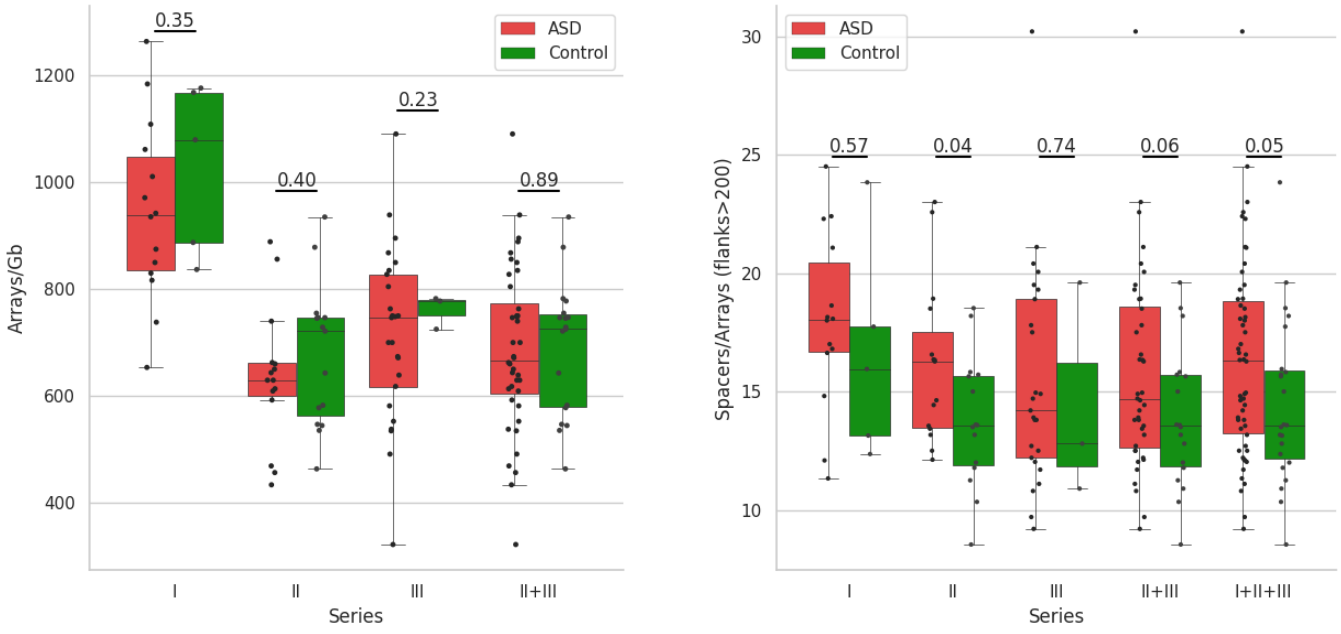
3. Results

3.1. Distribution of the CRISPR-Cas loci and its parameters

For our analysis we used publicly available siquencing data of the microbiomes for children with ASD and the contol group. In the original study faeces samples of 54 children with ASD and 23 healthy children aged 1 to 9 years old were sequenced in three independent series, differing in the sequencing protocols (Table 1) [40]. After appropriate processing of the sequencing data, we performed the prediction of CRISPR arrays using the procedure developed in [36] with CRISPRFinder replaced by its improved version CRISPRCasFinder [41]. The improvement of the procedure allowed us to predict not only the reliable list of CRISPR arrays, but also *cas* genes located nearby, where possible (Figure 1).

The number of the arrays normalized to the size of the microbiome assembly varies from 321.43 to 1264.29 arrays per Gb for individual metagenomes. Parameters of the array distribution for individual samples are illustrated in Supplementary table 1, Figure 2 and Table 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 January 2022 doi:10.20944/preprints202201.0296.v1



(**A**) (**B**)

Figure 2. Box plots of the arrays' numbers per Gb of assembly (A) and numbers of spacers per array (B) and the p-values of their comparisons for ASD and the control in all series and their allowed combinations (see Text). P-values are indicated above the pair of box plots for the compared datasets.

Table 2. Parameters of individual samples, their assemblies and arrays for healthy and ASD microbiomes for all series.

Parameters		Series I	Series II	Series III
ASD	Age	4.50±2.47	3.20±0.77	3.60±0.96
	Assembly size (Gb)	0.17±0.05	0.18±0.06	0.21±0.07
	Arrays	161.14±58.66	115.93±45.23	155.08±63.86
	Complete arrays (flanks>200)	22.21±10.24	22.00±9.58	33.04±15.34
	Arrays near *cas*	23.86±9.99	25.87±10.38	37.12±16.69
	Spacers	1247.57±488.94	895.07±321.10	1241.16±511.32
	Protospacers/Spacers (%)	6.09±2.06	6.73±2.42	4.64±1.77
Control	Age	3.40±0.55	2.87±0.52	3.0±0.0
	Assembly size (Gb)	0.16±0.02	0.18±0.04	0.23±0.05
	Arrays	160.60±16.62	122.13±31.64	172.33±38.53
	Complete arrays (flanks >200)	22.00±8.03	23.47±5.74	37.00±19.08
	Arrays near *cas*	24.80±8.84	26.07±5.90	39.00±13.45
	Spacers	1265.40±226.49	923.73±277.29	1381.67±465.66
	Protospacers/Spacers (%)	5.62±2.17	6.36±2.66	8.62±2.34

In order to compare the occurrence of CRISPR-Cas systems in individual microbiomes between children with ASD and the control group, we performed statistical analysis independently for each series. The reason for that is the sensitivity of the number of identified CRISPR arrays to sequencing and assembly parameters, particularly read

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 January 2022 doi:10.20944/preprints202201.0296.v1

length. Namely these parameters varied among the protocols for the data obtaining for the different series (Table 1). Thus, we could not analyse all the data simultaneously straightaway. Instead, we had to check the coincidence of the distribution expectations first. For this purpose, we used Welsch's adaptation of the Student's t-test after confirmation of normality for the distributions for each dataset with the Shapiro-Wilk test (see Data and Algorithms and Figure 2).

The comparison demonstrated smaller number of arrays in microbiomes of children with ASD, than in ones of the control group for all series. However, all the differences were insignificant even before the correction for multiple testing. To overcome this obstacle, we tried to combine the data of different series. But the pairwise comparison of arrays' numbers for ASD or the control in different series allowed joining only Series II and Series III, with none of them being allowed to join with Series I. However, the combined data demonstrated no significance in the test as well (Figure 2A).

Among other microbiome parameters tested for differences between children with ASD and the control group, only the length of complete arrays (i.e., average number of spacers in the arrays flanked with more than 200 nt in their contigs) demonstrated the similar results. In all comparisons, the complete arrays contained more spacers in the microbiomes of children with ASD (see Figure 2B). Here the differences were insignificant for all separate series except for the second one. In it, p-value was 0.04, i.e., under the significance level for single comparisons. Here, according to the statistical test, combining of all three series was allowed in pairs. For the combination of Series II and Series III datasets p-value was still higher than 0.05. But, after combining them with Series I datasets, it decreased to values lower than 0.05, the selected significance level for single comparisons. However, in all cases the differences remained to be insignificant because of the need for the correction for multiple testings.

All comparisons were rechecked using the Mann-Whitney-Wilcoxon test with exactly the same results.

3.2. Search for markers of the disease among CRISPR-Cas systems or their elements

The CRISPR-Cas systems represent a convinent platform for the method of close bacterial strains distinction [49]. That is why we considered their elements, direct repeats and particular spacers, as candidates for markers of the disease. To reveal such systems and elements, we selected arrays, which shared exact or very similar repeats, that were present in microbiomes of at least two children with ASD, but absent in microbiomes of the control group. For each series, such sets were respectively small and not numerous. For instance, for the Series II data, including equal numbers of ASD and healthy children's microbiomes, there were only three such sets with arrays from three different individual microbiomes each – one set with four arrays (two arrays from the same sample in the set) and two ones containing three arrays from different samples each. For different series the repeats of the distinguishing sets differed substantially, and there was no repeat distinguishing children with ASD from the control group at least in two series (data not shown).

Similarly, we failed to find particular spacers systematically distinguishing microbiomes of children with ASD. For example, for Series II there were no spacers represented in more than four individual microbiomes of children with ASD and not occurred in microbiomes of the control group. Moreover, even such spacers were different for the different series.

For the opposite task, as *Clostridium bolteae* was reported as a candidate marker species of the disease, because it was found in many microbiomes of the children with ASD [25], we checked our data for presence of the CRISPR-Cas system intrinsic for the species. *C. bolteae*, recently renamed into *Enterocloster bolteae* [50], is represented in current

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 January 2022 doi:10.20944/preprints202201.0296.v1

version of CRISPRCasDB by its two strains – ATCC BAA-613 and CBBP-2. Both strains include two CRISPR loci and one *cas* locus of type IC in their genomes. Repeats of the arrays adjacent to *cas* genes differ by one nucleotide between the strains and repeats of the distal arrays are identical. Both repeats are specific for the species, but the distal one is also present in two more species annotated as Lachnospiraceae bacterium or Lachnoclostridium sp. with one mismatch. We checked for the presence of these two repeats in all datasets. Both repeats occurred in the microbiomes of children with ASD more frequently. The sequences indentical to one of the copies of the repeat adjacent to the *cas* locus was found in twelve out of 54 individual samples of microbiomes for children with ASD in contrast to four out of 23 samples for the control group. The distal repeat was found in fifthteen samples out of 54 samples in the datasets with the ASD mark in contrast to only two out of 23 samples for the control group (Table 3). For the same dataset the lists of the individual microbiome samples contained these two repeats were different but overlapping in all cases. Thus, in our data, two CRISPR repeats of *C.boltae*, the adjacent repeat to the cas locus and the distal one, were present roughly one and one third or three times more frequently in the ASD samples than in the control group, correspondingly.

Table 3. Direct repeats from ATCC BAA-613 and CBBP-2 strains of Enterocloster bolteae and numbers of individual microbiomes, in which they were found.

Localisation in relation to cas	ATCC BAA-613 DR	CBBP-2 DR	Series I	Series II	Series III
			ASD/ Control	ASD/ Control	ASD/ Control
Adjacent to *cas*	GTCTCCGTCCTC GCGGGCGGAGT GGGTTGAAAT	ATTTCAACCCAC TCCGCCCACGAG GACGGAGAC	3/0	4/2	4/1
Distal from *cas*	ATTTCAATCCAC AAGGCTCTCGCG AGCCTCGAC	GTCGAGGCTCGC GAGAGCCTTGTG GATTGAAAT	3/0	4/2	8/0

In addition we analysed metagenomic contigs annotated as Clostridium bolteae or Enterocloster bolteae with MMseqs2 in order to check whether they could contain CRISPR arrays with other direct repeats. According to automatic annotation, each individual microbiome included one or more contigs assigned to the species for both children with ASD and the control groups. But there were no CRISPR arrays with any other direct repeat identified in the search.

3.3. Search for protospacers

We searched for protospacers in the genomic assemblies from the bacterial and viral sections of the RefSeq database. To distinguish protospacers from hits with spacers in CRISPR arrays, we had also performed a parallel comparison with the array repeats (see Data and Algorithms). As a result, the protospacers were found only for a minor fraction of the spacers in these databases (Supplementary table 1). The fraction varied from 1.7% to 31.2% and 1.9% to 36.4% for children with ASD and the control group, respectively.

We compared the fractions of spacers with protospacers in individual microbiomes between ASD and the control. According to the Welch's *t*-test, we had failed to reject the null hypothesis about the equality of expectations, i.e., the distributions of the value did not differ significantly. Moreover, the mean values of the fractions did not correlate among series (Table 2 and Supplementary table 1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 January 2022 doi:10.20944/preprints202201.0296.v1

We analysed the sets of bacteriophages containing protospacers. For this purpose, we compared the lists of phages providing protospacers for microbiomes of children with ASD and the control group from all our datasets. All lists were not numerous, with substantial intersections between lists for ASD and the control in each series (Supplementary table 2). For instance, for Series II the lists consisted of 56 phages for the case of ASD and 51 phages for the control group. The intersection of the lists of Series II included 21 organisms and, in particular, all the phages with more than four protospacers in any of the datasets (Supplementary table 2). The comparison of the taxonomical distributions of phages for the whole combined datasets for ASD and the control did not demonstrate substantial differences on the level of families (Figure 3).

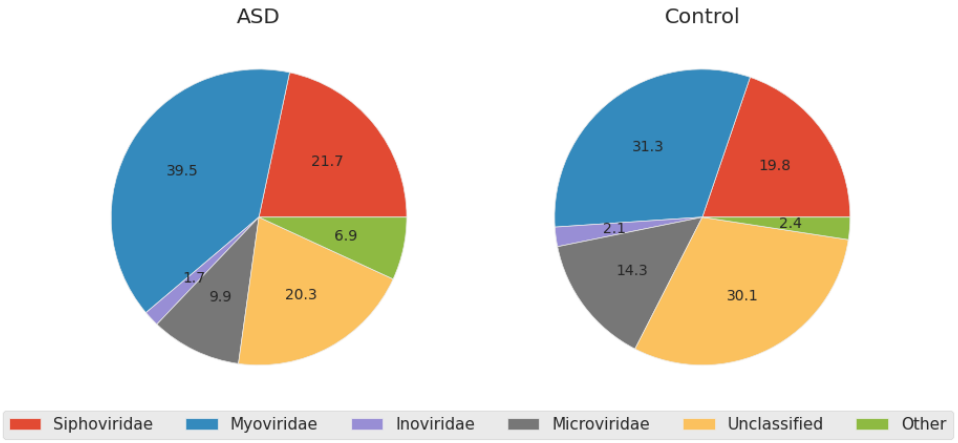


Figure 3. Piecharts for the distributions of the number of protospacers among different bacteriophage families for the combined datasets of microbiomes of children with ASD *and the control group.*

In order to analyse locations of multiple protospacers in the same known bacteria we performed the following procedure. We selected such bacterial Refseq records that harboured protospacers for more than any ten individual microbiomes. There were no requirements on the the specificity of the datasets here. For all three series a total of 26 such records were selected by this procedure (Supplementary table 3). Almost all protospacers in these bacteria were tightly grouped in their location. The prediction of the prophage regions with special tools and careful analysis of proteins encoded in these locations confirmed the bacteriophage origin of the corresponding chromosomal segments. Only two identified protospacers in two different bacteria had no signs of bacteriophage origin nearby (see Supplementary table 3, "NMTQ01000037.1" and "QSFJ01000016.1" sheats). There was no preference for microbiomes of children with ASD or the control group for records either in the number of individual microbiomes obtaining hits in the analysed bacterial genomes, or the total number of protospacers in each genome (data not shown).

4. Discussion

The problem of *a priori* CRISPR array identification has been solved quite efficiently for completed genomes [51] mainly owing to the rigid structural features of the arrays. But the same features provide additional difficulties for sequencing assembly. That is why this problem is much more complicated for massive highly fragmented data, like metagenomes, and is not closed yet. Computational tools designed for this purpose, like Met-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 January 2022 doi:10.20944/preprints202201.0296.v1

aCrast [52] or CRASS [53] , do not cover the area exhaustively. The former approach requires a list of predefined repeats as an input and, thus, is useless for the search of previously unknown arrays and respective systems. The latter one is strongly limited by the read length, and fails to restore the order of spacers automatically in the case of respectively short reads.

Alternative approach is provided by the use of a combination of tools, suitable for completed genomes. For example, for CRISPRCasMeta, an online service for systems identification in metagenomic data [54], main modules of CRISPRCasFinder are used in combination with the CRT program. In [55] a modified version of CRT, CRT-CLI [56], was used together with Piler-CR. Here, we used a slightly updated scheme based on all three original tools, CRISPRFinder, Piler-CR and CRT, suggested in [36].

For the distribution of CRISPR arrays in microbiomes of children with ASD and the control group, the only differences we observed were in the number of arrays per Gb of assembly and the number of spacers per array (Figure 2). These differences were slight and insignificant for our datasets, yet systematic. In our opinion, the main reason for the absence of significance is the small sizes of the datasets. We failed to overcome this obstacle by combining the data of different series. The data of different series were sequenced with different protocols and demonstrated significant variance in the numbers of array per Gb of assembly, in particular, for combinations of Series I and Series II or Series I and Series III, but not in the number of spacers in arrays. The step-by-step data combining for the latter parameter was accompanied with the decrease of p-value down to the value a bit lower than 0.05, the selected significance level for a single comparison. However, this observation remains insignificant after the correction for multiple comparisons.

As CRISPR-Cas systems are distributed more or less evenly among different taxa, the most probable explanation for the observed decrease of the number of arrays per Gb seems to be the general diversity reduction, which was demonstrated for the datasets we used [40]. The diversity reduction had also been observed in a number of other studies on the microbiomes of children with ASD [30,31]. However, as we already noticed in the introduction, some publications described the opposite effect [33].

Elongation of the arrays, on average, could reflect the bacteriophage burst in the ASD microbiomes, which, in turn, could be the reason for the diversity reduction. Unfortunately, almost nothing is known about the phage content of the gut microbiome in ASD. An alternative explanation for the array elongation for children with ASD could lay in the possibility of retention of the already used spacers for a bit longer due to the relaxation of the interspecies competition as a result of the diversity reduction.

We tried to find any markers for the disease among CRISPR-Cas systems or their elements, but failed. Neither repeats nor spacers distinguished microbiomes of children with ASD from the healthy ones: there were no systems' elements widely spread in ASD and not occurring in the control group. All candidate elements with biased distribution were found only in a minor fraction of the ASD microbiomes and were not reproduced among series. Thus, we believe that the bias was accidental in all these cases. Even systems of *Enterocloster bolteae*, the only particular species named in literature as characteristic for the ASD microbiomes, being quite specific in general, were found in ASD subjects only twice as often in the control group and were not widely spread. The most obvious reason for that, in our opinion, is the heterogeneity of the disease , as the diagnosis unifies different conditions with similar symptoms, but not etiology [57].

We expected that, according to the taxonomical changes and much more knowledge on the healthy gut microbiome, we would find notable differences in the comparison of protospacers for systems of children with ASD and the control group. That is why we compared fractions of spacers with protospacers in the known genomes and the sources of these protospacers. The formers differed neither significantly, nor systematically. Also, we failed to find substantial differences in the comparison of the lists of phages providing

 doi:10.20944/preprints202201.0296.v1

protospacers or their taxonomy. The phage lists intersected in their essential parts and all phages providing multiple protospacers belonged to the intersection. The overall number of the protospacers in phages of any family did not differ substantially between children with ASD and the control group as well. Thus, if phages do prevail in the microbiomes of children with ASD, as we suggested earlier in this section, this prevalence occurs rather due to the number of phages than to their diversity.

In the recent study on the oral microbiome, it was suggested that the CRISPR-Cas systems could participate in the interspecies competition [58] based on the comparison of the sources for protospacers of CRISPR arrays. In order to test this suggestion in the gut microbiome, both for children with ASD and the control group, we analysed the localization of the protospacers in bacterial genomes, providing them for multiple individual samples. It was demonstrated, that almost all protospacers were located in compact regions of chromosomal DNA, identified as prophage ones by special tools or based on the annotated function or best blast hits of the coded proteins. Only two protospacers, out of several hundreds checked, in two different bacteria were located separately from the others and not inside or close to the genes of potentially phage origin. Therefore, the most probable function of the microbiome protospacers is the antiphage defence, rather than participation in the interspecies competition. Thus, the considered phenomenon is at least not common in the gut microbiomes, both for children with ASD and for the healthy ones.

5. Conclusions

Here we analysed the CRISPR-Cas systems, their parameters and distribution in gut microbiota in health and disease. Their differences were demonstrated for the first time, to the best of our knowledge. Besides, the study provides the first indication of the involvement of bacteriophages in the changes of microbiome content in ASD.

According to our comparative analysis of the CRISPR-Cas systems, microbiomes of children with ASD contain fewer systems than the microbiomes of healthy children, but the arrays of these systems contain more spacers on average. The number of samples analysed here was insufficient and the observations were not significant but systematic. That is why the described patterns need to be tested additionally on the massive microbiome data.

Besides, the CRISPR arrays in microbiomes of children with ASD can be distinguished from the arrays of healthy children neither with the fractions of spacers with protospacers in known genomes, nor with the sets of known bacteriophages providing protospacers. Almost all bacterial protospacers of the gut microbiome systems for both children with ASD and the healthy ones are located in the prophage islands, that leaves no room for the systems to participate in the interspecies competition.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Supplementary table 1: Parameters of CRISPR arrays for all individual samples of all series, Supplementary table 2: Bacteriophages providing protospacers for the microbiome CRISPR-Cas systems and total number of protospacers from microbiomes of children with ASD and the control group, Supplementary table 3: Detailed description of the 26 bacterial loci harboring multiple protospacers. Each sheet corresponds to one bacterial RefSeq record, coloring reflects the algorithm with that the phage's origin of the protospacer location was identified.

Author Contributions: Conceptualization, Olga Averina, Valeriy Danilenko and Irena Artamonova; Methodology, Irena Artamonova; Software, Natalia Zakharevich and Mikhail Nikitin; Supervision, Irena Artamonova; Validation, Natalia Zakharevich, Mikhail Nikitin, Alexey Kovtun, Vsevolod Malov and Irena Artamonova; Writing – original draft, Mikhail Nikitin and Irena Artamonova; Writing – review & editing, Natalia Zakharevich, Alexey Kovtun, Olga Averina and Irena Artamonova. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Russian Foundation for Basic Research (RFBR) grant no. 18-29-07087 and partially supported by Russian program of Fundamental Research for State Academies №0112-2019-0001 and №0112-2019-0007.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Thaiss, C.A.; Zmora, N.; Levy, M.; Elinav, E. The microbiome and innate immunity. *Nature* **2016**, *535*, 65-74.
2. Visconti, A.; Le Roy, C.I.; Rosa, F.; Rossi, N.; Martin, T.C.; Mohney, R.P.; Li, W.; de Rinaldis, E.; Bell, J.T.; Venter, J.C., *et al.* Interplay between the human gut microbiome and host metabolism. *Nat Commun* **2019**, *10*, 4505.
3. Sampson, T.R.; Mazmanian, S.K. Control of brain development, function, and behavior by the microbiome. *Cell Host Microbe* **2015**, *17*, 565-576.
4. Hofer, U. The majority is uncultured. *Nat Rev Microbiol* **2018**, *16*, 716-717.
5. Specht, T.; Szymanski, M.; Barciszewska, M.Z.; Barciszewski, J.; Erdmann, V.A. Compilation of 5S rRNA and 5S rRNA gene sequences. *Nucleic Acids Res* **1997**, *25*, 96-97.
6. Patel, J.B. 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Mol Diagn* **2001**, *6*, 313-321.
7. Woo, P.C.; Lau, S.K.; Teng, J.L.; Tse, H.; Yuen, K.Y. Then and now: Use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clin Microbiol Infect* **2008**, *14*, 908-934.
8. Turnbaugh, P.J.; Ley, R.E.; Hamady, M.; Fraser-Liggett, C.M.; Knight, R.; Gordon, J.I. The human microbiome project. *Nature* **2007**, *449*, 804-810.
9. Ehrlich, S.D.; Consortium, M. Metahit: The european union project on metagenomics of the human intestinal tract. *Metagenomics of the Human Body* **2011**, 307-316.
10. Tyakht, A.V.; Alexeev, D.G.; Popenko, A.S.; Kostryukova, E.S.; Govorun, V.M. Rural and urban microbiota: To be or not to be? *Gut Microbes* **2014**, *5*, 351-356.
11. McDonald, D.; Hyde, E.; Debelius, J.W.; Morton, J.T.; Gonzalez, A.; Ackermann, G.; Aksenov, A.A.; Behsaz, B.; Brennan, C.; Chen, Y., *et al.* American gut: An open platform for citizen science microbiome research. *mSystems* **2018**, *3*.
12. Qin, J.; Li, Y.; Cai, Z.; Li, S.; Zhu, J.; Zhang, F.; Liang, S.; Zhang, W.; Guan, Y.; Shen, D., *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **2012**, *490*, 55-60.
13. Torres, J.; Mehandru, S.; Colombel, J.F.; Peyrin-Biroulet, L. Crohn's disease. *Lancet* **2017**, *389*, 1741-1755.
14. Guo, X.Y.; Liu, X.J.; Hao, J.Y. Gut microbiota in ulcerative colitis: Insights on pathogenesis and treatment. *J Dig Dis* **2020**, *21*, 147-159.
15. John, G.K.; Mullin, G.E. The gut microbiome and obesity. *Curr Oncol Rep* **2016**, *18*, 45.
16. Gupta, A.; Osadchiy, V.; Mayer, E.A. Brain-gut-microbiome interactions in obesity and food addiction. *Nat Rev Gastroenterol Hepatol* **2020**, *17*, 655-672.
17. Vogt, N.M.; Kerby, R.L.; Dill-McFarland, K.A.; Harding, S.J.; Merluzzi, A.P.; Johnson, S.C.; Carlsson, C.M.; Asthana, S.; Zetterberg, H.; Blennow, K., *et al.* Gut microbiome alterations in alzheimer's disease. *Sci Rep* **2017**, *7*, 13537.
18. Caputi, V.; Giron, M.C. Microbiome-gut-brain axis and toll-like receptors in parkinson's disease. *Int J Mol Sci* **2018**, *19*.
19. Wimberley, T.; Agerbo, E.; Pedersen, C.B.; Dalsgaard, S.; Horsdal, H.T.; Mortensen, P.B.; Thompson, W.K.; Kohler-Forsberg, O.; Yolken, R.H. Otitis media, antibiotics, and risk of autism spectrum disorder. *Autism Res* **2018**, *11*, 1432-1440.
20. Montecucco, C.; Schiavo, G. Mechanism of action of tetanus and botulinum neurotoxins. *Mol Microbiol* **1994**, *13*, 1-8.
21. Pardo, C.A.; Buckley, A.; Thurm, A.; Lee, L.C.; Azhagiri, A.; Neville, D.M.; Swedo, S.E. A pilot open-label trial of minocycline in patients with autism and regressive features. *J Neurodev Disord* **2013**, *5*, 9.

22. Finegold, S.M.; Molitoris, D.; Song, Y.; Liu, C.; Vaisanen, M.L.; Bolte, E.; McTeague, M.; Sandler, R.; Wexler, H.; Marlowe, E.M., *et al.* Gastrointestinal microflora studies in late-onset autism. *Clin Infect Dis* **2002**, *35*, S6-S16.
23. Sandler, R.H.; Finegold, S.M.; Bolte, E.R.; Buchanan, C.P.; Maxwell, A.P.; Vaisanen, M.L.; Nelson, M.N.; Wexler, H.M. Short-term benefit from oral vancomycin treatment of regressive-onset autism. *J Child Neurol* **2000**, *15*, 429-435.
24. Finegold, S.M.; Dowd, S.E.; Gontcharova, V.; Liu, C.; Henley, K.E.; Wolcott, R.D.; Youn, E.; Summanen, P.H.; Granpeesheh, D.; Dixon, D., *et al.* Pyrosequencing study of fecal microflora of autistic and control children. *Anaerobe* **2010**, *16*, 444-453.
25. Song, Y.L.; Liu, C.X.; Finegold, S.A. Real-time PCR quantitation of Clostridia in feces of autistic children. *Appl Environ Microb* **2004**, *70*, 6459-6465.
26. Tomova, A.; Soltys, K.; Repiska, G.; Palkova, L.; Filcikova, D.; Minarik, G.; Turna, J.; Prochotska, K.; Babinska, K.; Ostatnikova, D. Specificity of gut microbiota in children with autism spectrum disorder in slovakia and its correlation with astrocytes activity marker and specific behavioural patterns. *Physiol Behav* **2020**, *214*, 112745.
27. Zou, R.; Xu, F.; Wang, Y.; Duan, M.; Guo, M.; Zhang, Q.; Zhao, H.; Zheng, H. Changes in the gut microbiota of children with autism spectrum disorder. *Autism Res* **2020**, *13*, 1614-1625.
28. Adams, J.B.; Johansen, L.J.; Powell, L.D.; Quig, D.; Rubin, R.A. Gastrointestinal flora and gastrointestinal status in children with autism--comparisons to typical children and correlation with autism severity. *BMC Gastroenterol* **2011**, *11*, 22.
29. Kang, D.W.; Park, J.G.; Ilhan, Z.E.; Wallstrom, G.; Labaer, J.; Adams, J.B.; Krajmalnik-Brown, R. Reduced incidence of prevotella and other fermenters in intestinal microflora of autistic children. *PLoS One* **2013**, *8*, e68322.
30. Ma, B.; Liang, J.; Dai, M.; Wang, J.; Luo, J.; Zhang, Z.; Jing, J. Altered gut microbiota in chinese children with autism spectrum disorders. *Front Cell Infect Microbiol* **2019**, *9*, 40.
31. Kang, D.W.; Ilhan, Z.E.; Isern, N.G.; Hoyt, D.W.; Howsmon, D.P.; Shaffer, M.; Lozupone, C.A.; Hahn, J.; Adams, J.B.; Krajmalnik-Brown, R. Differences in fecal microbial metabolites and microbiota of children with autism spectrum disorders. *Anaerobe* **2018**, *49*, 121-131.
32. Gondalia, S.V.; Palombo, E.A.; Knowles, S.R.; Cox, S.B.; Meyer, D.; Austin, D.W. Molecular characterisation of gastrointestinal microbiota of children with autism (with and without gastrointestinal dysfunction) and their neurotypical siblings. *Autism Res* **2012**, *5*, 419-427.
33. Ding, X.; Xu, Y.; Zhang, X.; Zhang, L.; Duan, G.; Song, C.; Li, Z.; Yang, Y.; Wang, Y.; Wang, X., *et al.* Gut microbiota changes in patients with autism spectrum disorders. *J Psychiatr Res* **2020**, *129*, 149-159.
34. Nussenzweig, P.M.; Marraffini, L.A. Molecular mechanisms of CRISPR-Cas immunity in bacteria. *Annu Rev Genet* **2020**, *54*, 93-120.
35. McGinn, J.; Marraffini, L.A. Molecular mechanisms of c CRISPR-Cas spacer acquisition. *Nat Rev Microbiol* **2019**, *17*, 7-12.
36. Sorokin, V.A.; Gelfand, M.S.; Artamonova, II. Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. *Appl Environ Microbiol* **2010**, *76*, 2136-2144.
37. Burstein, D.; Harrington, L.B.; Strutt, S.C.; Probst, A.J.; Anantharaman, K.; Thomas, B.C.; Doudna, J.A.; Banfield, J.F. New CRISPR-Cas systems from uncultivated microbes. *Nature* **2017**, *542*, 237-241.
38. Munch, P.C.; Franzosa, E.A.; Stecher, B.; McHardy, A.C.; Huttenhower, C. Identification of natural CRISPR-Cas systems and targets in the human microbiome. *Cell Host Microbe* **2021**, *29*, 94-106 e104.
39. Gogleva, A.A.; Gelfand, M.S.; Artamonova, II. Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs. *BMC Genomics* **2014**, *15*, 202.
40. Averina, O.V.; Kovtun, A.S.; Polyakova, S.I.; Savilova, A.M.; Rebrikov, D.V.; Danilenko, V.N. The bacterial neurometabolic signature of the gut microbiota of young children with autism spectrum disorders. *J Med Microbiol* **2020**, *69*, 558-571.
41. Couvin, D.; Bernheim, A.; Toffano-Nioche, C.; Touchon, M.; Michalik, J.; Neron, B.; Rocha, E.P.C.; Vergnaud, G.; Gautheret, D.; Pourcel, C. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for cas proteins. *Nucleic Acids Res* **2018**, *46*, W246-W251.

42. Ghodsi, M.; Liu, B.; Pop, M. DNACLUST: Accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics* **2011**, *12*, 271.
43. Pourcel, C.; Touchon, M.; Villeriot, N.; Vernadet, J.P.; Couvin, D.; Toffano-Nioche, C.; Vergnaud, G. CRISPRCasDB a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res* **2020**, *48*, D535-D544.
44. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* **1997**, *25*, 3389-3402.
45. Steinegger, M.; Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **2017**, *35*, 1026-1028.
46. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D., *et al.* Reference sequence (RefSeq) database at ncbi: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **2016**, *44*, D733-745.
47. Song, W.; Sun, H.X.; Zhang, C.; Cheng, L.; Peng, Y.; Deng, Z.; Wang, D.; Wang, Y.; Hu, M.; Liu, W., *et al.* Prophage Hunter: An integrative hunting tool for active prophages. *Nucleic Acids Res* **2019**, *47*, W74-W80.
48. Zhou, Y.; Liang, Y.; Lynch, K.H.; Dennis, J.J.; Wishart, D.S. PHAST: A fast phage search tool. *Nucleic Acids Res* **2011**, *39*, W347-352.
49. Barrangou, R.; Horvath, P. CRISPR: New horizons in phage resistance and strain identification. *Annu Rev Food Sci Technol* **2012**, *3*, 143-162.
50. Haas, K.N.; Blanchard, J.L. Reclassification of the *Clostridium clostridioforme* and *Clostridium sphenoides* clades as *Enterocloster* gen. Nov. And *Lacrimispora* gen. Nov., including reclassification of 15 taxa. *Int J Syst Evol Microbiol* **2020**, *70*, 23-34.
51. Alkhnbashi, O.S.; Meier, T.; Mitrofanov, A.; Backofen, R.; Voss, B. CRISPR-Cas bioinformatics. *Methods* **2020**, *172*, 3-11.
52. Moller, A.G.; Liang, C. Metacrast: Reference-guided extraction of CRISPR spacers from unassembled metagenomes. *PeerJ* **2017**, *5*, e3788.
53. Skennerton, C.T.; Imelfort, M.; Tyson, G.W. CRASS: Identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res* **2013**, *41*, e105.
54. CRISPRCasMeta. Available online at https://crisprcas.i2bc.paris-saclay.fr/CrisprCasMeta/Index
55. Pavlova, Y.S.; Paez-Espino, D.; Morozov, A.Y.; Belalov, I.S. Searching for fat tails in CRISPR-Cas systems: Data analysis and mathematical modeling. *PLoS Comput Biol* **2021**, *17*, e1008841.
56. Huntemann, M.; Ivanova, N.N.; Mavromatis, K.; Tripp, H.J.; Paez-Espino, D.; Palaniappan, K.; Szeto, E.; Pillay, M.; Chen, I.M.; Pati, A., *et al.* The standard operating procedure of the doe-jgi microbial genome annotation pipeline (mgap v.4). *Stand Genomic Sci* **2015**, *10*, 86.
57. Lai, M.C.; Lombardo, M.V.; Chakrabarti, B.; Baron-Cohen, S. Subgrouping the autism "spectrum": Reflections on dsm-5. *PLoS Biol* **2013**, *11*, e1001544.
58. Gong, T.; Zeng, J.; Tang, B.; Zhou, X.; Li, Y. CRISPR-Cas systems in oral microbiome: From immune defense to physiological regulation. *Mol Oral Microbiol* **2020**, *35*, 41-48.