Computational Prediction and Analysis of CRISPR-Cas systems in Chinese Human Gut Metagenomic Samples

Tatiana Cabral Mangericão[1,2]

CRISPR has been becoming a hot topic as a power technique for genome editing for human and other higher organisms. The original CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats coupled with CRISPR-associated proteins) is an important adaptive defence system for prokaryotes that provides resistance against invading elements such as viruses and plasmids. A CRISPR cassette contains short nucleotide sequences called spacers. These unique regions retain a history of the interactions between prokaryotes and their invaders in individual strains and ecosystems.

One important ecosystem in the human body is the human gut, a rich habitat populated by a great diversity of microorganisms. Metagenome sequencing has been widely applied for studying the gut microbiome. Most efforts in metagenomic studies have been focused on profiling taxa compositions and gene catalogues, and identifying their associations with human health. Less attention has been paid to the analysis of the ecosystems of microbiomes themselves, especially their CRISPR composition.

In this work, assembled human metagenomic gut samples of Chinese diabetic, and healthy, individuals, was investigated for its CRISPR content, focusing on all important elements of a CRISPR system: the repeat sequences, the spacers and the *cas* genes.

Keywords— CRISPR; Cas genes; Human Gut; Microbiome; Metagenome

I. Introduction

The human body is host to this complex community of symbiotic, pathogenic and commensal microorganisms (microbiome), whose abundance is estimated to exceed the number of human cells by at least an order of **[1].** It has been estimated that the microbes in our bodies collectively make up to 100 trillion cells, tenfold the number of human cells **[2].** To understand and exploit the impact of the microbes on human health and wellbeing, it is necessary to interpret the content, diversity and function of the microbial community.

Metagenomic sequencing has been proven to be a powerful tool for analysing complex microbial communities. And in this context many projects have been developed all over the world to study human microbiomes of multiple body sites. For example, the US-based Human Microbiome Project (HMP) **[3]** and the EU-based MetaHIT project **[4]** have generated resources that can enable the comprehensive characterization of the human microbiome and analysis of its role in human health and disease**.**

Among all body sites, the diversity of microorganisms in the human gut is known to be among the highest **[1]**. This community has been discovered to be associated with human physiology through processes related to development, nutrition, immunity, and resistance to pathogens **[5-9].** While bacteria are responsible for these functions and most research has focused on the interaction of bacteria with the host, bacteriophages, in turn, influence the composition and abundances of bacteria in the human gut **[10,11].**

Bacteria and Archaea have evolved defence and regulatory mechanisms to cope with various environmental stressors, especially virus attacks. The understanding on this arsenal has been expanded by the discovery of the versatile CRISPR-Cas system. Bacteria can remember their viral invaders by sampling short DNA sequences, known as protospacers, from the genetic materials of viruses or phage. These sequences become integrated into the bacterium's own DNA, specifically into an array of repeat sequences called clustered regularly interspaced short palindromic repeats (CRISPR). The integrated sequences are called spacers **[12]**. When these sequences are transcribed and processed into small RNAs, they guide a multifunctional protein complex (Cas proteins – CRISPR associated proteins) to recognize and cleave incoming foreign genetic material **[13]**. The diversity of Cas genes suggests that multiple pathways have been developed to use the basic information contained in the CRISPR cassettes in diverse defence mechanisms **[14]**. This adaptive immunity system was first observed in Escherichia coli in 1987, although its significance was not straightaway apparent. Since then, CRISPR arrays have been identified in approximately 40% of Bacteria and 90% of Archaea **[15]**.

CRISPR cassettes were already characterized across human body sites in different individuals in independent projects **[16-18]** and as a part of the Human Microbiome Project (HMP) **[14]** with particular focus in the gut metagenome **[14,19,20].**

[1] Department of Bioengineering, Instituto Superior Técnico (IST), Lisbon, Portugal

[2] MOE Key Lab Division,TNLIST/Center for Synthetic and Systems Biology, and Department of Automation, Tsinghua University, Beijing 100084, China

e-mail: tatiana.mangericao@tecnico.ulisboa.pt.

II. MATERIALS AND METHODS

A. Metagenome Dataset

The metagenomic set used in this project was downloaded from a cluster, containing metagenomic samples, from the MOE Key Lab of Bioinformatics/Bioinformatics Division in Tsinghua University. The metagenomic data comes from a Metagenome-Wide Association Study (MWAS) intended at identifying associations between gut microbiota and Type-2 Diabetes **[21]**. The sequenced and analysed data from the study can be freely accessed online at GigaScience database (GigaDB).

Bacterial DNA was extracted from faecal samples and sequenced by whole-genome shotgun (WGS) using the Illumina GAIIx and HiSeq2000. The DNA samples were collected from 145 Chinese Han individuals (from the age between 14 and 59). The dataset used for this project corresponds to the Stage I sequenced samples in the study and is composed by samples of two major groups: 71 diabetic individuals - classified DLF, DLM, DOF and DOM and 74 non-diabetic individuals, used as controls - NLF, NLM, NOF and NOM **[21]**. For further reference, we name the diabetic samples dataset as T2D+ and the controls as T2D-.

Using SOAPdenovo, the individual metagenomes were assembled in contigs, with the average size each of 10.687 bp. The whole dataset size comprised 15.96 Gb (16,345Mb) and a total number of 8,039,994 contigs.

More information about the individuals (sex, age, BMI…) originating the metagenomic samples can be consulted at **Supplemental File 1**.

B. Identification and analysis of CRISPR cassettes

To construct a set of CRISPR cassettes for the metagenomic dataset two freely available CRISPR-finding algorithms were used, PILER-CR **[22]** and CRISPR Recognition Tool (CRT) **[23]**, together with a simple filtering procedure.

Both algorithms were downloaded and run in Mac OSX's Terminal environment using a simple command-line interface. For PILER-CR the command line is *./pilercr –in <input_file> -out <report_file_name>* and for CRT, *java –cp CRT1.2-CLI.jar crt <input _file>*. Either algorithm only accepts an input file in FASTA format.

For both algorithms the default parameters were used:

Table 1- Most relevant default parameters used for both CRISPR detection algorithms; The repeat range is the length within a repeat sequence must fall, minimum length and maximum length; Spacer range is the minimum and maximum length a spacer sequence must have to be accepted by both algorithms; Min. repeats in a cassette is the minimum number of repeats that a CRISPR cassette must have to be considered valid.

Algorithm	Repeat Range	Spacer Range	Min. repeats in a cassette
PILER-CR	16 - 64	8 - 64	3
CRT	19 - 38	19 - 48	3

The filtering procedure applied to the output obtained from both softwares has one simple step: compare the resulting collection of CRISPR arrays from PILER-CR and CRT, and keep only the CRISPRs which were predicted by both programs simultaneously, meaning they shared the same repeat consensus sequence and spacers. The resulting collection of CRISPR arrays is considered a sufficient reliable collection.

With both PILER-CR and CRT we have access to the repeat consensus sequence and the spacers' sequences. There is also information about the CRISPR array length, and position, in the contig.

C. Repeat clustering and analysis

From the collection of reliable CRISPR arrays we extracted the repeat consensus sequences.

For a more extensive analysis we choose to only include CRISPRs belonging to metagenomic samples originating from T2D diabetic patients (T2D+).

The first step of repeat analysis is the construction of a set of unique repeat sequences. This was made manually with the aid of Microsoft Excel's conditional formatting tool, which allows for a rapid identification of duplicated sequences.

An ID was attributed to all resulting repeats, making the distinction between unique repeats and redundant repeat sequences. The first were assigned with a number ID (example: >1), and the latter, who are present in more that one CRISPR cassette, assigned with a letter ID (example: >AA), so in further analysis only the unique sequence is considered.

Each unique repeat sequence was matched against the CRISPR database, CRISPRdb (accessible at http://crispr.u-psud.fr/crispr/BLAST/CRISPRsBlast.php) **[24]**, using a standard BLAST with an *e-value* threshold of 0,01, in order to find hits for known repeats in our set and to access the possibility of new repeat sequences. For each repeat that had a match within the database, we attributed a preliminary taxonomic label to the corresponding CRISPR array, indicating the related bacterial strain and NCBI identification.

All repeats from the non-redundant set were analysed with the CRISPRmap tool **[25]**. The webserver from which this tool

is accessible is located online at: http://rna.informatik.uni-freiburg.de/CRISPRmap/.

The input field included only the CRISPR repeat sequences, in FASTA format, properly identified with unique ID's. Optimization of reading direction of input sequences is done in the data processing. Even if this option weren't checked in the webserver interface, CRISPRmap would still check both directions of the given input sequences for their occurrence in their CRISPR repeats database. In order for an occurrence to be reported, there has to be a 100% match to one of the consensus repeat sequences. CRISPRmap version used is v2.1.3-2014, containing 4719 consensus repeats covering 24 sequence families and 18 structural motifs.

D. Taxonomy of metagenomic contigs containing CRISPR cassettes

To determine contig taxonomy, contigs were subjected to a BLASTX (version 2.2.32) search against the non-redundant protein collection **[26]**, which includes all non-redundant GenBank coding sequences translations, Protein Databank (PDB), SwissProt, PIR and PRF databases, excluding environmental samples from whole-genome shotgun (WGS) projects. This database was last updated in November 2015 and encompasses 74,367,285 protein sequences.

Contig query sequences were inputted, in FASTA format, in the BLASTX platform of NCBI, accessible at: http://blast.ncbi.nlm.nih.gov/blast/Blast.cgi. Parameters used corresponded to the default parameters devised for the algorithm: the scoring matrix is BLOSOM62, with a cost to create and extend a gap in the alignment of 11 and 1, respectively. The maximum number of aligned sequences displayed is 100, with the attention that the actual number of alignments might be greater than this. The length of the seed that initiates an alignment, word size, is 6. The box related to Filters and Masking of low complexity regions is checked.

The *e-value* threshold isn't a changeable parameter, but when analysing the output, matches with an e-value larger than 1e-6 are automatically dismissed.

Taxonomic labels were assigned manually based on the degree of consistency with the taxonomy origin of the top hits and the taxonomic BLAST report. The latter summarizes the BLAST output classification and the relationships between all of the organisms found in the BLAST hit list. The taxonomic label was assigned at a phylum, class, family and genus level, when possible. If not, the contig was assigned with a nonspecific taxonomic label, "Bacteria".

A contig might not been assigned a taxonomy for a number of reasons: the CRISPR cassette covers (almost) the entire length of the contig; the flanking regions of the cassette contains only universal Cas genes, which phylogeny does not necessarily reflect taxonomy due to the frequent phenomena of horizontal gene transfer; there is no significant similarity to any entry in the query database.

With BLASTX it is also possible to identify Cas genes. Whenever hits with description fields containing the words "cas" and "crispr" appeared in the output, these were collected for further manual analysis. Again, hits above the *e-value* threshold of $1e^{-6}$ were not considered. For some outputs, the hit list contained a significant number of hits for universal Cas genes, Cas1 and Cas2, indicating also the associated CRISPR type and subtype. In these cases, it was possible to assign a classification to the associated CRISPR-Cas system. In other cases, type was only assigned taking into account the signature Cas genes of each type.

All hits were also confirmed with the CRISPRFinder **[27]** option to extract the flanking sequences of the submitted contigs containing CRISPR cassettes, and search for similarities between these upstream and downstream regions in a local cas bank database, *casdb*, using the BLASTX algorithm.

E. Identification of Protospacers

In order to identify and label spacer origin, CRISPRTarget program was used **[28].** This software works with a BLASTN algorithm, but differentiates itself from the NCBI associated algorithm by the default parameters used. The CRISPRTarget BLASTN parameters favour gapless matches but allow a number of mismatches at this screening stage, with a higher gap penalty 10, rather than 5 than the NCBI defaults. The mismatch penalty is -1 and the *e-value* filter is 1. Also, there is also no filter or masking for low complexity.

The software query sequences included only the spacer sequences, in FASTA format, extracted from the set of reliable CRISPR cassettes. Redundant spacers are removed during the query processing and listed in a separate file, which is later used to create the collection of non-redundant spacers.

Target databases for the Target BLASTN search included the default GenBank-Phage and RefSeq plasmid. The first is one of the smallest of the GenBank divisions containing 6,800 sequences with 88 million bases. The latter, related to RefSeq databases (reference sequences of NCBI), contains 3,707 sequences with a total of 282 million bases.

The output of the program was filtered using the cutoff score and number of mismatches. Default value for cutoff is 20, but only matches with a score equal, or higher, than 25 were contemplated. Also, matches with more than 3 mismatches were discarded.

To infer about the presence of a largely common bacteriophage, Gut phage BED-2012 or crAssphage **[29]**, a BLAST N alignment was used between all the non-redundant spacer sequences and the complete genome sequence of the phage. The phage has the GenBank accession number JQ995537.1 and a total length of 97065 bp.

III. RESULTS AND DISCUSSION

In this section we present the results obtained, following the designed and detailed pipeline, in methods, and our view on them.

A. Metagenomic dataset

The metagenomic dataset used for this project was selected for its availability, body site location and geographic origin. The data originated from faecal samples (gut samples) belonging to individuals from a geographic background yet to be studied and characterized for its CRISPR content. The individuals are naturals from China.

The dataset used is publicly available (see Materials and Methods) and was constructed as part as a metagenome-wide association study of gut microbiota in type-2 diabetes **[12]**. The individuals are Chinese type-2 diabetic and non-diabetic individuals. Type-2 Diabetes (T2D) is a complex disorder influenced by both genetic and environmental components, and has become a major public health issue across the world. Although the focus of this project was not on trying to associate CRISPR with this disease in particular, the base for further work in this field is laid down with this work.

It should be noted that the total size of the dataset analysed didn't match the original one from the study, since 3 samples were missing: DLM022, DOF007 and NLM032. So, in total, we analysed 142 samples instead of 145.

B. Detection and characterization of CRISPR cassettes

We used two publicly available CRISPR-detecting tools, CRT and PILER-CR, to search for CRISPR cassettes in human whole-metagenome assemblies from Chinese diabetic and non-diabetic individuals (T2D+ and T2D-, respectively).

The choice of the programs fell on its wide acceptance by the scientific community, as they are two of the most used softwares for identification of *de novo* CRISPRs in genomic data. As well as in its user interface simplicity and proved performance: they are both fast, memory efficient, and provide high levels of quality, precision and recall **[23]**. Even so, they are not perfect. For example, CRT has some unreliability problems since instead of reading the input as a series of contigs, it considers contigs of each individual as connected to each other as a unique uninterrupted sequence, making no difference between them, which is an incorrect assumption. To surpass this problem, we didn't consider for analysis predicted CRISPR that existed in a range that included more than one contig.

In 2012, Mina Rho and colleagues modified CRT to consider incomplete repeats at the ends of contigs from whole-metagenome assembly, and called the new algorithm metaCRT **[14]**. However, although there's an online platform to access this modified version of CRT (http://omictools.com/metacrt-s10717.html), the link is broken, rendering the program unusable for the time being of this project.

Regarding PILER-CR, within one array, with default parameters, the algorithm allows a fair amount of variability in the repeat and spacer's length, in order to maximize sensitivity. This may allow identification of inactive ("fossil") arrays, and may in rare cases also induce false positives due to other classes of repeats such as microsatellites, Long Terminal repeats and arrays of RNA genes.

The use of only two of the three existent, and most used, CRISPR detection tools was due to the fact that CRISPR Finder is only accessible as an online tool and it's not available for download. Trying to upload large data to the website made the processing time too long for large volume data like the one we were using (whole-metagenomic assemblies). And to add to this, the web server only allows input sequences up to 67,000,000 bp, which is not enough for some of the samples in our dataset (for example, sample DOM001 and DOM010 are 77,165,163 bp and 86,503,610 bp long, respectively). Nevertheless, in further steps, it is applicable to smaller data like individual contigs.

There is also a fairly recent detection tool named CRASS **[30]**. This algorithm was purposely made for identifying CRISPR in shotgun metagenomic data from Illumina, Ion Torrent PGM, Roche 454, and Sanger platforms, using an iterative search approach that does not rely on preassembled contigs or prior knowledge Meaning, contrary to what happens with both PILER-CR and CRT, it searches through raw metagenomic data (reads) for direct repeat (DR)-containing reads and reconstructs the CRISPR loci. CRASS algorithm was showed to provide a fast running time, high specificity and sensitivity when identifying CRISPR DRs. It's strict filtering steps, required to correctly group individual reads into DR types, sometimes might result in missing spacers from CRISPR loci where the DR sequence was not highly conserved **[30].**

In this present project, we could not use CRASS, since we didn't have access to the shotgun reads from the Chinese individual's gut metagenomes. For further field related-projects, applying CRASS to the raw data and complement this output with the ones from the programs used could provide us some further insight over the CRISPR content existent in the samples. The extra sensitivity and specificity of Crass might reveal more details about population heterogeneity and phage-host interactions, which would not have been discovered in assembled data. Even more, it could help complete the set of CRISPR cassettes, with ones produced by reads that weren't assembled into contigs during the assembly process.

With PILER-CR and CRT we found a total of 3,022 and 3,110 candidate CRISPR cassettes, respectively. Among the total number of CRISPRs found by PILER-CR, 1,563 cassettes belong to the 73 individuals from the control group and the remaining 1,459 cassettes belong to the 69 individuals with type-2 diabetes. In what concerns CRT, 1,601 cassettes belong to the control group and the remaining 1,509 cassettes belong to the type-2 diabetes samples. With both algorithms, CRISPR cassettes have been predicted in all single individuals.

After applying the filtering step, we have a resulting collection of 1,325 CRISPR cassettes, from which 630 belong to the T2D+ (type-2 diabetes positive) dataset and 695 to the T2D- (control group) dataset. To these CRISPR cassettes correspond a total of 25,879 spacers and 1,325 repeats. The

total number of spacers predicted for the CRISPR cassettes is 8,140 for the diabetic dataset and 17,739 for the control group. See **Supplementary file 2**.

To exclude false predictions of CRISPR cassettes in the metagenomic data, a basic filtering procedure was applied. This procedure, made only of one simple step, consists in comparing the results obtained from both programs and retaining the cassettes predicted simultaneously by the two programs. These cassettes share the same repeat sequence (with no mismatch allowed) and the same set of spacers. Both programs have its flaws and limitations and may output some false CRISPR predictions, so, in applying this filtering step, we add more reliability to the collection of resulting CRISPR cassettes.

To the remaining predicted CRISPR cassettes, we simply excluded them from further analysis and didn't apply any further filtering.

In 2014, for a study in comparative analysis of CRISPR cassettes from public human gut metagenomes, Gogleva and colleagues **[31]** devised a three step filtering system to form a reliable set of CRISPR cassettes: (1) Cassettes predicted by more than two CRISPR predicting programs, should be kept. Using more than one algorithm is recommended as they have complementary strengths for precision and recall; (2) Cassettes predicted by less than two programs, i.e. candidate cassettes, which are adjacent to Cas genes, are added to the set of cassettes constructed in the first step; (3) candidate cassettes whose repeat consensus sequence is part of a repeat cluster containing repeats from already established as reliable CRISPR cassettes, are added to the set. Comparing our process to theirs it is possible to observe that we only applied the first step, which we consider to be the most important. We accept that we might lose some putative CRISPR cassettes by not following the whole pipeline, but the simple comparison of results is considered to be reliable enough.

Comparing both outputs from the used algorithms it is possible to perceive that the number of overlapping cassettes is somewhat low. On average, only 40% of the cassettes are matched. This could be explained by the fact that we are searching for 100% identical repeats between the software's outputs, and sometimes it exists a single nucleotide mismatch, or a couple of extra nucleotides in the repeat sequence, that compromises the overlapping.

In other cases it might happen that the number of spacers is not equal, differentiating by more than two spacers, even if the repeat sequence is matched. Usually, when such thing happens it's because either the repeat closer to the leader sequence, or the one further from it, present some mismatches with the consensus repeat sequence, and CRT might include it in the CRISPR cassette, while PILER- CR doesn't, causing a failed match between cassettes (See Figure 9).

On the other hand, it is possible to observe that both softwares produce similar results, meaning, they usually predict the same number of CRISPR cassettes for each individual. Which may lead to affirm that they are in fact similar in performance.

It should be referred that if we analyse both healthy and diabetic individuals separately the results are very similar, meaning that we couldn't find an obvious distinction between the number of CRISPR cassettes or number of spacers found. Thus, for further analysis, we chose to focus on the diabetic group (T2D+).

1) CRISPR database, CRISPRdb

The non-redundant set of repeat sequences (set of 362 repeat sequences) was submitted to a BLAST search against the CRISPR database (see Materials and Methods).

The CRISPR database comprises a collection of known CRISPR cassettes from publically available bacterial and archaeal genomes. The database base was last updated in August 2015 and has 1,176 bacterial strains with convincing CRISPR(s), from 2,612 bacterial analysed genomes, and 126 out of 150 Archaea genomes with convincing CRISPR(s).

In total, 271 sequences had a match in the database (approximately 75%) with an e-value smaller than 0,01, leaving 91 unmatched sequences. The latter, that weren't matched to CRISPRdb, can possibly be repeats belonging to novel CRISPR cassettes that haven't been reported before.

In what concerns the matched repeats, by using only an e-value cut-off, it's possible that one part of a repeat partially matches part of one or more direct repeats in the database. But the two repeats could be quite different at other positions. This may introduce some false positive hits (that is, the 271 hits are not all true matches). In fact, the addition of this step to the pipeline only works for doing a preliminary taxonomy of the CRISPR cassette and origin contig, through the repeat sequence. To attribute a final, more reliable taxonomy, the use of other complementary tools, like BLAST X, is recommended. This step also allows discovering known and novel CRISPR repeats.

In total, a number of 119 different strains were found. The majority of the matches belonged to the genus Clostridium and Eubacterium, with 31 and 18, out of 271 matches, respectively. Even so, they only represent 11% and 7% of the results, which ends up being not particularly relevant if we look at the "bigger picture", contributing to the idea that the microbial population in the human gut is in fact very complex and diverse. The most represented species was *Megamonas hypermegale* ART12/1 (15 hits from 271). The second most matched species, within our repeats, were *Eubacterium rectale* (11 hits from 271) and *Faecalibacterium prausnitzii* (9 hits from 271), all species commonly found in the human gut microbiota **[32]**. The last two species are bacteria that are part of the colonic microbiomes responsible for the fermentation of hexose and pentose sugars **[33]**.

As discussed, the results obtained may not be completely accurate, but they indicated a large diversity of CRISPR-containing strains present in the human gut microbiome.

The detailed results are given in **Supplemental File 4**. For the set of significant clusters, defined as redundant repeat sequences that contain six or more recurrences in different CRISPR cassettes, see **Supplemental File 5**.

2) Taxonomy of CRISPR containing contigs

To define the taxonomic origins of contigs containing the identified cassettes, a BLASTX-based procedure was used (see Materials and Methods for more details). It should be noted that this procedure was only applied to contigs

containing either unique CRISPR repeats or contigs containing repeats from the set of "significant clusters", totalling 317 analysed contigs (234 contigs with unique repeats plus 83 contigs harbouring the most represented repeats).

The short length of some metagenomic contigs combined with the propensity of Cas genes to horizontal gene transfer makes taxonomic predictions for CRISPR-containing contigs somewhat difficult. However, it was possible to assign a taxonomy label, at least at the domain level, to 256 of 317 cassettes (approx. 80%). For approx. 20% of the contigs, no strong evidence of any particular phylum was detected, so these contigs were generically assigned to "Bacteria".

The largest fraction of contigs with assigned taxonomy belonged to Firmicutes. 134 contigs of this origin were observed (approx. 57% of total contigs), with the majority of them belonging to the Clostridia (96 contigs – 72%) and Negativicutes (26 contigs – 19 %) classes.

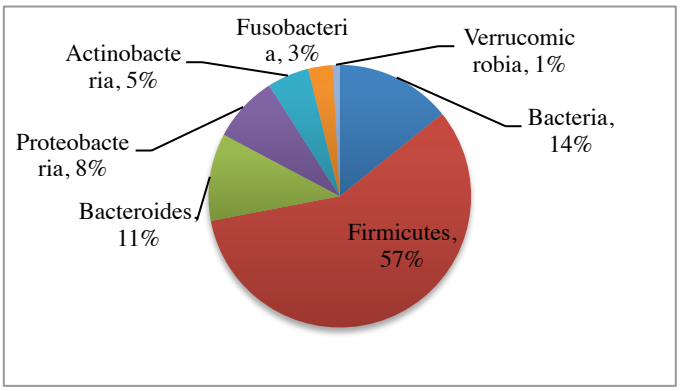

Figure 1 - Taxonomy of CRISPR-containing contigs with unique repeats; Firmicutes has the largest fraction attributed, counting for more than half of the analysed contigs; Bacteria is referent to the fraction of contigs to which wasn't possible to attribute a taxonomy, so it was attributed this generic label.

The second major group, in the analysed dataset, comprised 25 contigs from the Bacteroidetes phyla (approx.19 %), being all assigned to the Bacteroidales order. *Prevotella* and *Bacteroides* dominated the genus assignment

The human gut microbiome is vast, and consists of about 10^{14} bacterial cells, which is ten times the number of cells in the human body. Of the plus fifty known phyla, most of the human microbiota is composed by less than ten (and mostly six) phyla. Bacteria from other phyla, usually of plant origin, that may be present in skin, nasopharyngeal, or gut samples; are generally infrequent (<0.01% of the sequences) and probably represent transient carriage from food- and air-borne exposures **[34]**.

Based on 16S rRNA-based surveys and by direct sequencing of genetic material, it is apparent that in general the adult gut is a complex community dominated by two bacterial phyla, Firmicutes and Bacteroidetes (comprising approximately 90% of the bacterial ecosystem), with other phyla including Actinobacteria, Proteobacteria, Verrucomicrobia and Fusobacteria, being present in lower proportions. Greater variations exist below the phylum level, although certain butyrate-producing bacteria, including *Faecalibacterium prausnitzii*, *Roseburia intestinalis* and *Bacteroides funiformis*, have been identified as key members of the adult gut microbiota **[34,35].**

The taxonomic labelling assigned, when possible, to the CRISPR-containing contigs, confirms what was expected relatively to the phyla dominance of Firmicutes and Bacteroides.

In the T2D metagenome-wide association study by Qin, J. and colleagues **[21]**, they investigated the subpopulations of the individual samples, and identified three enterotypes in the Chinese samples. A principal component analysis (PCA) revealed that to these enterotypes corresponded to several highly abundant genera, including *Bacteroides, Prevotella, Bifidobacterium* and *Ruminococcus.* In fact, representatives of these genuses were found within our CRISPR-containing contigs.

Assigning a taxonomic label to these contigs, serves to confirm the literature, presenting a large diversity at a genus level, reflecting the human gut microbial variety, and also to associate CRISPR cassettes to specific species.

A functional CRISPR-*cas* immune system consists of both a CRISPR cassette and *cas* genes **[36]**. It was attributed, where possible, a classification to the identified systems according to repeat types and associated *cas* genes. The latter were found in flanking sequences of 58 from the 234 cassettes analysed, correspondent to the unique repeats. In a considerable fraction of flanking sequences the only identified *cas* genes were *cas* 1 and/or *cas* 2, universal markers of most CRISPR-*cas* systems, therefore not applicable for differentiating between system types.

Among the cassettes that could be classified at type or associated subtype-level, according to the characteristic *cas* genes, 27 cassettes where assigned to CRISPR-*cas* type I; 15 cassettes to type II; 13 to type III and 3 to putative new type V. For 40 cassettes it was possible to also assign a subtype. For more detail results see **Supplemental File 6**.

As it is possible to verify, not all cassettes had associated *cas* genes. This could have happened due to the short length of the contig containing the said cassette, being the latter covered almost entirely by the repeat-spacer block. In other cases, the finding of only one or two *cas* genes, may be due to a incomplete CRISPR-*cas* loci, which usually happens in about 12% of the bacterial genomes **[37]**. Complete single-unit loci are most commonly type I systems, whereas putative type V systems are rare (<2% overall). The latter is in fact apparent with the results obtained where type V systems represent 5% of the identified systems.

The most abundant CRISPR-Cas system was subtype I-C, representing 32% of the total sample, followed by subtype III-A (7 cassettes, 17%) and subtype II-A (5 cassettes, 12%). Subtype I-B and I-E had a similar distribution (4 cassettes, 10%).

Different bacterial phyla usually show distinct trends in the distribution of CRISPR-Cas systems **[37]**. According to a recent review from Makarova and colleagues, the phylum Firmicutes, one of the most represented in the human gut, accounts for most of the subtype II-A systems. It is not possible to confirm this since only 5 cassettes were assigned to

subtype II-A. Nevertheless, 3 of them did belong to Firmicutes. Most of contigs assigned to this phyla were identified has being from type I systems, the majority belonging to subtype I-C, with no representatives from subtype I-E. The latter is usually strongly associated with Actinobacteria. In fact, from the 4 CRISPR-Cas systems assigned to this subtype, two belonged to this phylum and the other two to Proteobacteria. As expected, Proteobacteria lacked subtype I-A.

Considering the enormous importance of type II systems in biotechnology, it is important to refer that this type was significantly represented in two phyla: Firmicutes and Proteobacteria (87%). Type II systems, constituted by a single subunit crRNA-effector module, dramatically differ from types I and III, being, by far, the simplest in term of number of genes. The main player is the *cas* 9 gene (also appearing in the nomenclature has *csn1* and *csx12*) which encodes the multi domain protein complex responsible by the expression and interference phases in the CRISPR immunity. All three subtypes, II-A, B and C, are very similar, only differing in one gene. Subtype II-A systems include an additional gene, *csn2,* which is considered a signature gene for this subtype. Actually, it is possible to verify, that whenever this gene was found, the subtype was automatically assigned, even if this was the only gene identifiable in the CRISPR-containing contig.

It should be noted that some Cas genes sometimes might appear associated with a certain bacterium phylum that is not usually the carrier of this type of genes. This may happen due to horizontal transfer, which Cas genes are frequently subjected to.

3) Reconstruction of a CRISPR-Cas array

The analysis of CRISPR cassettes wouldn't be complete without the reconstruction of a CRISPR-Cas locus.

The latter is constituted by the array containing the short direct repeats separated by short variable DNA sequences, the spacers, adjacent to a leader sequence, and flanked by diverse Cas genes involved in the CRISPR adaptive immunity.
For this purpose we choose one contig where it was identified a CRISPR cassette and all associated Cas genes. This contig contains a system representative of type I.

Type I systems are defined by the presence of a multisubunit crRNA-effector complex, the Cascade complex. All the CRISPR-Cas loci from this type include a signature cas3 gene (or its variant cas3') **[31]** Type I systems are currently divided into seven subtypes, I-A to I-F and I-U. Each subtype has a defined combination of signature genes and distinct features of operon organization. Type I-C, which was the most represented among the CRISPR-Cas systems analysed, is a derivative from subtype I-B, descendant of the ancestral type I gene arrangement (cas1-cas2-cas3-cas4-cas5-cas6-cas7-cas8) However, subtype I-C lacks Cas6, which seems to be functionally replaced by Cas5.

In contig 'scaffold28217_8', with a length of 8,260 bp, from individual DLM019, was found a complete type I-C CRISPR-Cas system. Within the contig was possible to annotate the gene cas2, ranging from position 1118 to 1336, followed by Cas1 (1351 to 2367bp), Cas4 (2451 to 2924bp), Cas7 (3047 to 3877bp), Cas8c (3943 to 5652bp), Cas5 (5832 to 6437bp) and finally, the signature type I gene cas3, with a range between position 7019 and 1993. The corresponding CRISPR array is located upstream of the Cas genes, and directly adjacent to a leader sequence. The array is formed by a set of 9 repeat sequences (length equal to 33bp), interspaced by 8 spacers. The CRISPR array is harboured by a bacterium belonging to Firmicutes phylum, which, according to the top hits resulting form the BLASTX search, corresponds to *Clostridium sp*..

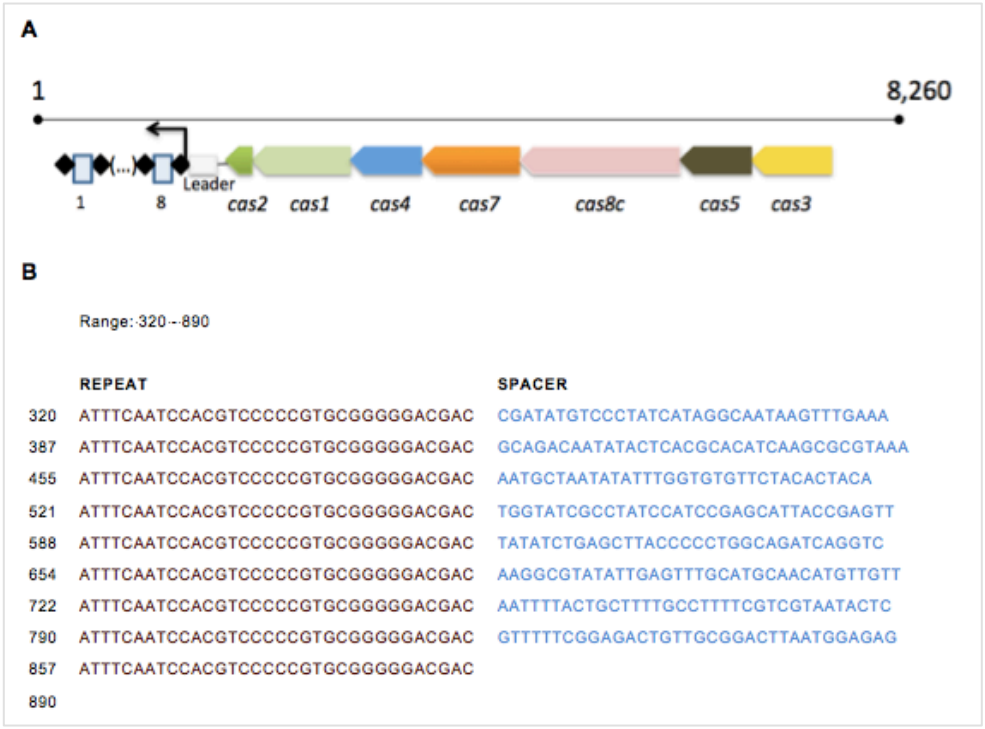

Figure 2 - Schematics of the architecture of an identified type I-C CRISPR-Cas system, from contig 'scaffold28217_8, originated in individual DLM019; Contig has a total size of 8,260 bp; (**A**) The annotated cas genes, downstream of the array, are constituted by the signature cas3 gene, characteristic of type I systems. Next is the cas5 gene, trailed by Cas8c and Cas7, completing the effector module. It is then followed by the core Cas genes, Cas4, Cas1 and Cas2; (**B**) the CRISPR array, with length 570bp, is composed of 8 spacers (blue coloured rectangles) and 9 repeat sequences (◆), with an adjacent leader sequence downstream of the first repeat; Spacer 1 is the one furthest from the leader sequence, so it is the first one inserted to the array. Spacer 4 is the newest; Note that the scheme is made on scale and does not represent the exact size of the genes or CRISPR array.

C. Identification and analysis of protospacers

Our set comprised 8145 spacer sequences, from which, 748, were shared by two or more CRISPR cassettes (corresponding to 351 sequences). Thus, the final non-redundant set had 7748 spacers.

The totality of the nr-spacer set was analysed with CRISPRTarget. This program was chosen for its specificity, since it was created with the sole purpose of matching CRISPR spacers with their protospacer pairs, searching in all available viral libraries. Practically, it is similar as using BLASTN from NCBI, but with different parameters.

Program results yielded 37 spacer-protospacer pairs. However, the output was filtered to only include pairs with a final score higher than 25 and 3 or less mismatches. See **Supplemental File 8** for the complete overview of the output of CRISPRTarget software.

The detected spacer-protospacers pairs corresponded to 21 different spacers, from which 13 matches were found in viral genomes of phages infecting *Enterobacteria, Bacteroides* and *Lactobacillus.*

More than one spacer had different matching protospacers with the same number of mismatches and score, but, in general, the protospacers taxonomy accorded in the target bacterial genus. Notably, one of these spacers had protospacers, all with two mismatches, in four different *Lactobacillus* phages: *Lactobacillus* phage Ld3, phage Ld27, phage Ld25A and phage c5. Even though they all have different denominations, these protospacers are known for infecting *Lactobacillus delbrueckii* subsp. *bulgaricus,* and display high levels of sequence identity to each other. These Lactobacilli phages belong to one of the most prevalent virus order, *Caudovirales,* and possess a long noncontractile tail, typical of the *Siphoviridae* family **[38]**. All the protospacer sequences occur in a region codifying a putative terminase large subunit.

Two different spacers, originating from different CRISPR arrays with similar repeat sequences, matched *Lactobacillus* phages phiLdb and Ld3, and three different spacers, also coming from arrays with the same repeat, matched protospacers in the *Bacteroides* phage φB124-14. The latter is a human gut-specific bacteriophage **[39]**. φB124-14 infects only a subset of closely related gut-associated *Bacteroides fragilis* strains. The protospacers occurred in coding sequences assigned to a hypothetical protein and to two putative capsid proteins, similar to major protein 2 and 3 (MP2, B40-8039; MP3, B40-8040) from *Bacteroides* phage B40-8. This phage belongs to the *Caudovirales* order and *Siphoviridae* family.

Eleven spacers matched protospacers residing in plasmids belonging to *Klebsiella pneumoniae* and *Escherichia coli,* from enterobacterial origin, *Bifidobacterium breve, Bacteroides fragilis, Lactobacillus spp.* and *Campylobacter spp.*. None of the plasmids matched relevant proteins.

The taxonomy of CRISPR-containing metagenomic contigs can be determined relying on either flanking sequences, as we've seen, or protospacers.

The virus specificity for targeting certain bacterial species, allows for it to be used as a taxonomic labelling tool. From the seven contigs containing spacers with a protospacer pair belonging to a phage genome, to only one was attributed a taxonomy using this approach, since to the others it had already been applied a label using BLAST X tool in the CRISPR array flanking sequences. The former was contig 'C392804_1' from sample DOF008, which was labelled at a genus level because of its matching protospacer from *Lactobacillus* phage AQ113 (from the Firmicutes phyla).

Comparing the protospacer taxonomy to the one from the contigs, four taxonomic labels demonstrated a good concordance, as the assignments agreed at least on the level of phyla. In particular, contig 'scaffold22656_1' from sample DOF008, had been assigned, as expected, to the *Lactobacillus* genus (Firmicutes phyla), and more specifically, to the *delbrueckii* subsp. *bulgaricus.*

Recently, there was the discovery of a previously unidentified bacteriophage present in the majority of published human faecal metagenomes, which was referred to as crAssphage **[29]**. Its ~97 kbp genome is one of the most abundant in publicly available metagenomes, comprising up to 90% and 22% of all reads in virus-like particle (VLP)-derived metagenomes and total community metagenomes, respectively; and it totals 1.68% of all human faecal metagenomic sequencing reads in the public databases **[29]**. To assert about the presence of crAssphage we used BLAST N to align all our spacers, from the nr set, with the complete genome sequence of the phage. Interestingly enough, contrary to what might be expected, the result only yielded one match.

The almost inexistent number of matches could possibly be explained by the fact that the potential virus sequences analysed are limited to the spacer sequences present in the predicted CRISPR cassettes. This means that viral sequences not associated with CRISPR aren't accounted for in the analysis, resulting in "lost information", which wouldn't happen if we were analysing a human gut virome. Other important point that could explain the number of matches is the fact that the sequences available from databases account for a low percentage of the estimated number of genomic sequences of the existing prokaryotic viruses **[40]**.

Bacteriophage associated with the human gut microbiome are likely to have an important impact on community structure and function, and provide a wealth of biotechnological opportunities. Despite this, knowledge of the ecology and composition of bacteriophage in the gut bacterial community remains poor, with few well-characterized gut-associated phage genomes currently available.

D. Comparative analysis with other human gut metagenomic datasets

After analysing the non-redundant sets of repeat and spacer sequences, a comparative analysis was conducted, focusing on both the repeats and spacers. For more detailed results see **Supplemental File 9**.

It should be noted that, for this project, the objective was not to ultimately make a full comparative analysis between the Chinese dataset and other human gut datasets, since, there is

no base to imply that the comparison would reflect differences between populations, as an analysis on how well each dataset can represent the population from which they were sampled, wasn't made. The differences can be because of the sampling, and can also be attributed to technical methods used in the different labs, e.g. the assembly algorithm and sequencing technology, or other factors.

1) Repeat set

In 2014, Gogleva and colleagues **[31]** published a study that used human gut metagenomic data from three open projects in order to characterize the CRISPR composition and dynamics of human-associated microbiota. The datasets included the Human Microbiome Project (HMP) gut samples, Distal Gut metagenome project (DG) samples and the assembled metagenomic datasets from healthy 13 Japanese individuals (JPN). The CRISPR detection software used included PILER-CR, CRT and CRISPRFinder.

The most identical dataset to the one used in this project, was the HMP dataset, comprising samples from 124 European adults of various ages (18-69) sequenced by Illumina GA machines. In this dataset used by Gogleva et al, the total number of contigs was 1,889,651, and the total length of the contigs comprised 3,732Mb. In the T2D data, both healthy and diabetic individuals, the total number of contigs is 8,039,994, and the total length of contigs comprised 15.96 Gb (16,345Mb) **[31]**

It should be noted that the data used by Gogleva and colleagues, referring to the HMP dataset, was downloaded from the website http://public.genomics.org.cn/BGI/gutmeta/UniSet/ (ref.41 in Gogleva et al, 2014) **[31]**). From this web address, as well as the basic information provided in their paper, it is lead to believe that the data used in the study was not HMP data, but instead, the MetaHIT data produced by Qin and colleagues. This, however, does not affect their results or ours.

Analysing briefly the results they obtained, many more CRISPRs are found in the Gogleva JPN dataset than the "HMP" data, even though the number of individuals and the number of contigs is much smaller. The JPN data is very small (only 463MB vs. 8.8 GB in the diabetic dataset) compared to the number of predicted CRISPRs it resulted in: 283 cassettes detected by both PILER-CR and CRT). If we simply extrapolate by size alone, we could predict 283*(8.8GB/463MB) = 5849 predicted CRISPRs in the diabetes data, which in reality didn't happen (630 CRISPR detected by both algorithms).

Comparing their obtained CRISPR repeats against our set of non-redundant repeats (362 sequences), it is possible to perceive some overlaps, mostly with the JPN data. Indeed, it was found a match for 49 identical repeat sequences, against 5 from "HMP" and 3 from DG. Matches between repeat sequences mostly signify a possible match in the CRISPR-containing contig, indicating that the CRISPR cassette belongs to the same bacterial genus or strain, even if there is no match with the spacer's sequences.

The highest number of matches to JPN could be explained thanks to the larger number of CRISPRs predicted for the latter, compared with the ones predicted for DG and "HMP" (13 and 61 cassettes, respectively); or, maybe, the closest geographical location from the individuals originating the samples (similar gut flora). In fact, data size is one factor, but the number of CRISPRs detected can be conditioned by a series of known or unknown factors, such as the type of bacteria present in the microbiota of the individual (only ~50% of bacteria has a CRISPR system), the size of the CRISPR system, which can dependent on the history of HT (horizontal transfer) and former invasions by phages, the program used for CRISPR detection, protocols and settings in the sequencing experiment, the quality of the sequencing data and assembly of the metagenomic sample. In the end, it's challenging to compare results and draw meaningful conclusions, as the grand truth is unknown in real data. This is an important question and deserves deeper investigation on both simulated data and real data for the future.

2) Spacer set

For a comparative analysis regarding the CRISPR spacers, another dataset was chosen, since Gogleva et al publicated results didn't include this data.

The CRISPR set identified by Stern and colleagues in raw MetaHIT reads contains 52,267 spacers, 48,484 of which are unique **[19]**. In this study they obtained the set of spacers by extracting them directly from the raw reads, and using them as probes to search for phage genomic segments within the assembled sequences of the metagenomes. This allowed them to identify and characterize a large catalogue of phages and other mobile elements, along with associated bacterial hosts, invading the human gut of European individuals.

Comparing these spacers with the spacer set identified in the scope of this project, comprising 8,145 spacers (with 7,748 unique sequences), only 41 matches were found, originating in 59 different cassettes. The matched spacers cover 5% of our unique spacers in the non-redundant set, but none of the matches was an identified spacer with an attributed taxonomy.

A couple of possibilities could explain such a low overlap between the two sets. First of all, the size of the data is considerably different, since the methods from which both sets were obtained vary in a significant way. In this study we are analysing the assembled data directly, so it's possible that we might miss some CRISPR cassettes due to the fact that reads containing putative spacers were not assembled in the contigs. Secondly, we are comparing two datasets originating in different geographically located populations, which are, from the outset, different in composition, because of a series of environmental and dietary factors. This affects the diversity of microorganisms present, as well as the diversity of invading phages, meaning, ultimately, that the individuals were not exposed to the same infecting viruses. Still, it's worth mentioning that both hypotheses require further experimental exploration.

IV. Conclusions and Future Perspectives

We analysed CRISPR content in a human gut metagenomic dataset of Chinese individuals of healthy and type-2 diabetes

groups, with a bigger emphasis on the latter. With some relatively newly released tools for CRISPR identification and post-processing, we were able to profile CRISPR cassettes and their corresponding repeats and spacers in the gut microbiome of this sample set. Comparison with the existing database show to some extent that the majority of the identified CRISPRs have been reported in the literature, while there is a quarter of the identifications that indicate newly discovered CRISPRs that have not been reported before.

The human gut microbiota is one of the most complicated microbial ecosystems in the human body and has important associations with human health. CRISPR is a major system that microbes use to deal with phage invasions and challenges. Therefore analysing the CRISPR composition of a microbiome and of a group of microbiomes can be very informative for understanding the history and function of the microbiome. Also, the microbiota is composed of highly diverse bacteria species that cannot be cultured. Identifying new CRISPRs from metagenome data might also provide an efficient approach for finding possible novel CRISPRs that may be used for genome editing applications.

The collection of spacers and repeats constructed in this work constitutes a base for further studies, and aids to complement the already existent inventories of CRISPR loci in human microbiomes. As multiple gut metagenome datasets have been published in the projects like HMP, MetaHIT and other projects that focus on specific groups of people or specific human diseases, it'll be interesting and promising to conduct more comparative studies among different datasets, which may lead to better understanding of the forming, shaping, changing and function of gut microbiome populations in different individuals.

Whole-metagenome assemblies are useful for identifying novel CRISPRs, with detection softwares like CRT and PILER-CR. Nonetheless, some CRISPR cassettes might be missed by whole-metagenome assembly, thus, for further studies, it is suggested to complement this methodology with a read based approach, using a targeted assembly approach, aiming for a better assembly of the CRISPR loci with more complete structures.

More advanced studies could explore more the Cas proteins and their relationship with CRISPR array repeats, for better understanding of the defence mechanism. Methods like CRISPRmap are important for further studies regarding CRISPR-Cas proteins systems. We should also pay more attention to the spacer sequences and, for example, on discovering how this defence mechanism recognizes short sequence motifs, known to be adjacent to the spacer precursor in the invaders genomes. Other suggestion is using the spacer organization to trace the viral exposure of the hosts.

Regarding the type-2 diabetes disease, no association with CRISPR-Cas systems was completed. The link between this complex disease and this adaptive immune system of bacteria could be further explored, in order to comprehend this disease impact on the human gut flora and CRISPR systems.

Appendix

All Supplemental Files are in the form of excel files.

Supplemental File 1 Information of the metagenome samples; The Chinese dataset consisted in faecal samples belonging to 145 Chinese individuals living in the south of China, collected by Shenzhen Second People's Hospital, Peking University Shenzhen Hospital and Medical Research Center of Guangdong General Hospital. The set comprised 71 type-2 diabetic individuals - classified DLF, DLM, DOF and DOM - and 74 healthy controls - NLF, NLM, NOF and NOM.

Supplemental file 2 This file shows the detailed info about the predicted CRISPR cassettes from both datasets, with PILER-CR and CRT softwares, after applying the filtering step. Present in sheet (A) is information about the number of cassettes found in each individual, cassette origin scaffold, length (in base pairs), number of spacers and repeat consensus sequences. In the same excel sheet there is statistical information about the outputs from both algorithms and the overlap between them. In sheet (B) is the collection of spacers referent to each CRISPR cassette, identified individually according to the individual metagenome they belong to.

Supplemental File 3 The file contains two sheets. Sheet (A) corresponds to the complete set of CRISPR repeats prevenient from the set of reliable cassettes relative to the diabetic dataset. Information includes repeat origin (sample ID and contig), as well as the repeat consensus sequence and attributed ID. Also, it is shown the collection of repeat clusters (with each respective sequence), with an alphabetic ID from A to DY; Sheet (B) contains the non-redundant collection of repeat sequences.

Supplemental File 4 Repeat sequence's BLAST hits against CRISPRdb database; The results for the BLAST search against CRISPRdb of known repeats include a table showing for each unique repeat consensus the corresponding hit in the database, strain name and NCBI code, as well as the e-value. The excel file also includes a summary of the results (percentage of unique repeats that found a hit in the database), as well as the distribution of the known repeats for the different Bacteria phylum. A colour code is applied for the most represented phylum.

Supplemental File 5 Information about the composition of the significant clusters collected from the set of repeat sequences. Significant clusters are shown in detail with reference to the CRISPRdb hit. Each cluster contains information about the corresponding repeat sequence and ID, as well as information about the contigs where the repeat is present. BLASTX and local cas bank output is also provided in the table.

Supplemental File 6 Detailed information about the set of unique repeats and the results from the BLASTX search against the set of non-redundant proteins of GenBank.

Supplemental File 7 CRISPRmap software output relative to the set of non-redundant repeats. The results for CRISPRmap contain the distribution of the unique repeat sequences for the 6 superclasses (A-F), sequence families and structural motifs. There is a summary table of the results as well as a detailed table with data about the sequence families. The file is composed of 6 different excel sheets. Sheet (A) comprises the unfiltered output of CRISPRmap software; Sheet (B) contains the CRISPRmap tree updated with our inputted repeats; Sheet (C) focuses on the Superclasses; Sheet (D) on the sequence families; Sheet (E) on the structural motifs and sheet (F) contains the CRISPR-Cas system type attribution according to CRISPRmap, as well as a taxonomy comparison between that output and BLASTX search output for the unique repeats. In sheets (C) and (D), some further details are presented for the sequence families and structural motifs, respectively.

Supplemental File 8 Collection of spacers for the Type-2 diabetic dataset and results from CRISPRTarget software for protospacer taxonomy. The collection of spacer sequences is presented in detail in this file: sheet (A) is for metagenomes from the group DLF; sheet (B) is for spacers from set DLM; sheet (C) for DOF and sheet (D) is for DOM individuals. In each sheet are the CRISPRTarget results regarding the spacers from the corresponding set. Sheet (E) comprises all the CRISPRTarget results and a comparison to the previously attributed taxonomy to the contigs containing the spacer sequences.

Supplemental File 9 Comparative analysis relative to the set of predicted repeats and spacers, with other set predicted for different metagenomic datasets.

V. Bibliography

1. Li, K., Bihan, M., Yooseph, S., and Methé, B. A. "Analyses of the Microbial Diversity across the Human Microbiome". *PLoS ONE* 7, no. 6 (2012): e32118

2. Qin, J. *et al*. "A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing". *Nature* 464 (2010): 59-65.

3. The Human Microbiome Project (HMP) [http://hmpdacc.org/overview/about.php].

4. The MetaHit project. [http://www.metahit.eu/index.php?id=453].

5. Conly, J. M., Stein, K., Worobetz, L., Rutledge-Harding, S. "The contribution of vitamin K2 (menaquinones) produced by the intestinal microflora to human nutritional requirements for vitamin K." *The American Journal of Gastroenterology* 89, , no. 6 (1994): 915-923.

6. Hill, M. J. "Intestinal Flora and Endogenous Vitamin Synthesis." *European Journal of Cancer Prevention* 6, Suppl 1 (1997):S43-5.

7. Cummings, J. H. "Microbial Digestion of Complex Carbohydrates in Man." *Proceedings of the Nutrition Society* 43, no. 1 (1984): 35-44.

8. Black, D. D., Jianhui, D., and H. Wang "Regulation of Apolipoprotein Secretion by Long-Chain Polyunsaturated Fatty Acids in Newborn Swine Intestinal Epithelial Cells." *Pediatric Research* 43 (1998): 98.

9. Backhed, F. "Host-Bacterial Mutualism in the Human Intestine." *Science* 307, no. 5717 (2005): 1915-920.

10. Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A. J., Thomson, N. R., Quail, M., Smith, F., Walker, D., Libberton, B., Fenton, A., Hall, N., Brockhurst, M. A. "Antagonistic coevolution accelerates molecular evolution." *Nature* 464, no. 7286 (2010): 275–278.

11. Riley, P. A. "Bacteriophages in autoimmune disease and other inflammatory conditions." *Medical Hypotheses* 62, no. 4 (2004): 493–498.

12. Yosef, I. and U. Qimron "Microbiology: How Bacteria Get Spacers from Invaders." *Nature* 519, no. 7542 (2015): 166-67.

13. Bhaya, D., Davison, M. and R. Barrangou. "CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation." *Annual Review of Genetics* 45, no.1 (2011): 273-97.

14. Rho, M., Wu Y. W., Tang, H., Doak, T. G. and Y. Ye. "Diverse CRISPRs Evolving in Human Microbiomes." *PLoS Genetics* 8, no. 6 (2012): 1-9.

15. Sorek, R., Kunin, V, and P. Hugenholtz "CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea." *Nature Reviews Microbioligy* 6 (2008): 181–186.

16. Pride, D.T, Sun, C. L., Salzman, J., Rao, N., Loomer, P., Armitage, G. C., Banfield, J. F. and D. A. Relman "Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time." *Genome Research* 21, no. 1 (2011): 126–136.

17. Pride, D.T., Salzman ,J., Relman, D. A. "Comparisons of clustered regularly interspaced short palindromic repeats and viromes in human saliva reveal bacterial adaptations to salivary viruses." *Environmental Microbiology* 14, no. 9 (2012): 2564–2576.

18. Robles-Sikisaka, R., Ly, M., Boehm, T., Naidu, M., Salzman, J. and D. T. Pride "Association between living environment and human oral viral ecology." *The ISME Journal* 7, no. 9 (2013): 1710–1724.

19. Stern, A., Mick, E., Tirosh, I., Sagy, O. and R. Sorek "CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome." *Genome Research* 22, no. 10 (2012): 1985–1994.

20. Mick, E., Stern, A. and R. Sorek "Holding a grudge: persisting anti-phage CRISPR immunity in multiple human gut microbiomes." *RNA Biology* 10, no.5 (2013): 900–906.

21. Qin, J., et al. "A metagenome-wide association study of gut microbiota in type 2 diabetes." *Nature* 490 (2012): 55-60.

22. Edgar, R. C. “PILER-CR: fast and accurate identification of CRISPR repeats.” *BMC Bioinformatics* 8, no. 18 (2007).

23. Bland , C., et al. "CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats." *BMC Bioinformatics* 8, no. 1 (2007): 209.

24. Grissa, I., Vergnaud, G. and C. Pourcel “The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats” *BMC Bioinformatics* 8 (2007): 172.

25. Lange, S. J., O. S. Alkhnbashi, D. Rose, S. Will, and R. Backofen. “CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems.” *Nucleic Acids Research* 41, no. 17 (2013): 8034-44

26. Benson, D. A., et al. “GenBank.” *Nucleic Acids Research* 41, no. (Database issue) (2013): D36–D42

27. Grissa, I., G. Vergnaud, and C. Pourcel. “CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats.” *Nucleic Acids Research* 35, no. Web Server Issue (2007): W52–W57.

28. Biswas, A., J. N. Gagnon, S. J. J. Brouns, and C. M. Brown. “CRISPRTarget Bioinformatic prediction and analysis of crRNA targets.” *RNA biology* 10, no. 5 (2013): 817–27.

29. Dutilh, B. E., et al. "A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes." *Nature Communications* 5 (2014): 4498.

30. Skennerton, C. T., M. Imelfort, and G. W. Tyson. “Crass: Identification and reconstruction of CRISPR from unassembled metagenomic data.” *Nucleic Acids Research* 41, no. 10 (2013): 105.

31. Gogleva, A. A., M. S. Gelfand, and I. I. Artamonova. “Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs.” *BMC Genomics* 15, no. 202 (2014): 1-15.

32. Sun, J., and E. Chang. “Exploring gut microbes in human health and disease: Pushing the envelope.” *Genes & Diseases* 1, no. 2 (2014): 132-9

33. Russell, W. R., L. Hoyles, H. J. Flint, and M. E. Dumas. “Colonic bacterial metabolites and human health.” *Current Opinion in Microbiology* 16 (2013): 246-54.

34.Cho, I., and M. J. Blaser. “The Human Microbiome: at the interface of health and disease.” *Nature Reviews Genetics* 13, no. 4 (2012): 260-70.

35. Tremaroli, V., and F. Bäckhed. “Functional interactions between the gut microbiota and host metabolism.” *Nature* 489 (2012): 242-9.

36. Makarova, K. S., Wolf, Y. I. and E. V. Koonin. "The basic building blocks and evolution of CRISPR–Cas systems." *Biochemical Society Transactions* 41 (2013): 1392–1400.

37. Makarova, K. S., et al. "An updated evolutionary classification of CRISPR-Cas systems." *Microbiology* 13 (2015): 1-15.

38. Casey, E., Mahony, J., O'Connell-Motherway, M., Bottacini, F., Cornelissen, A., Neve, H., Heller, K. J., Noben, J. P., Dal Bello, F., van Sinderen, D. “Molecular Characterization of Three Lactobacillus delbrueckii subsp. bulgaricus Phages.” *Applied and Environmental Microbiology* 80, no. 18 (2014): 5623–35.

39. Ogilvie, L. A., Caplin, J., Dedi, C., Diston, D., Cheek, E., Bowler, L., Taylor, H., Ebdon, J., and B. V. Jones “Comparative (meta)genomic analysis and ecological profiling of human gut-specific bacteriophage φB124-14.” *PLos One* 7, no. 4 (2012): e35053.

40. Rohwer, F. “Global phage diversity.” *Cell* 113 (2003): 171-82.