TÉCNICO LISBOA

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

Tatiana Cabral Mangericão

Thesis to obtain the Master of Science Degree in

Biological Engineering

Supervisors: Prof. Dr. Xuegong Zhang
Prof. Miguel Nobre Parreira Cacho Teixeira

Examination Committee

Chairperson: Prof. Arsénio do Carmo Sales Mendes Fialho
Supervisor: Prof. Miguel Nobre Parreira Cacho Teixeira
Member of the Committee: Prof. Jorge Humberto Gomes Leitão

November 2015

Acknowledgements

First, a word of appreciation and a big thank you to Dr. Xuegong Zhang, bioinformatics division director in Tsinghua University, for all his guidance and insight during the length of this project.

I would like to acknowledge undergraduate student Zhanhao Peng and PhD Hongfei Cui, from Tsinghua University, for all the computational analysis assistance . And to Yanhui Xu, also from Tsinghua University, for all the help provided during the beginng phase of this project.

Finally, a special thank you to my family and friends for their unconditional support.

謝謝! Thank you! Obrigada!

Abstract

Parallel to the eukaryotic immune system, bacteria harbor CRISPRs, a lineup of DNA direct repeats and spacers to promote immunity against invading mobile genetic elements, like phages. A CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) cassette, found within the genome of an organism, together with the adjacent Cas genes, form the CRISPR-Cas immune system.

The phage-host interactions involving CRISPRs have been studied in experiments with selected bacterial species and in completely sequenced genomes. However, these studies do not allow one to take bacterial population diversity and phage-host interaction dynamics into account. Using metagenomic data can fill this gap.

We conducted an analysis of CRISPR content in Human Gut metagenomic samples of Chinese individuals with type-2 diabetes and their healthy controls. Applying two available CRISPR-identification algorithms, PILER-CR and CRT, together with a simple filtering step, we identified 1,325 reliable CRISPR cassettes in the data. A richer analysis was made on the diabetic dataset, focusing on the repeats, spacers and cas genes. The whole constructed set of 362 unique repeat sequences was compared with known CRISPRs in the database CRISPRdb. A total of 271 repeats had matches in the database, and the remaining 91 repeats are potentially novel ones. CRISPR repeats were also submitted to a comprehensive clustering and classification using the web server tool CRISPRmap.

Further, only 37 of the 7,748 identified non-redundant spacers were coupled with protospacers from phage and plasmid databases.

Overral, the computational analysis of CRISPR composition, in metagenome sequencing data, is feasible. Providing an efficient approach for finding potential novel CRISPR arrays and for analysing the ecosystem and history of human microbiomes.

Keywords

CRISPR; Cas genes; Human Gut; Microbiome; Metagenome

Resumo

Paralelemente ao sistema de defesa eucariota, as bactérias abrigam os CRISPRs. Estes tratam-se de um vector de DNA com repetições directas e espaçadores, que promovem a imunidade contra a invasão de elementos genéticos móveis, como fagos. *Clustered Regularly Interspaced Short Palindrome Repeats*, CRISPR, encontrado no interior do genoma de um organismo, em conjunto com os genes adjacentes, *Cas*, formam o sistema imunitário CRISPR-Cas.

As interacções invasor-hospedeiro envolvendo o sistema CRISPR foram estudados em experiências com espécies bacterianas seleccionadas e em genomas completamente sequenciados. No entanto, estes estudos não permitem aferir sobre a diversidade populacional e dinâmica de interação bacteriana que existe em diversos ecossistemas. A sequenciação metagenómica veio preencher esta lacuna.

Foi conduzida uma análise ao conteúdo CRISPR em amostras metagenómicas do intestino humano de indivíduos chineses com diabetes tipo 2 e dos seus controles saudáveis. Aplicando dois algoritmos de identificação disponíveis, PILERr-CR e CRT, em conjunto com uma etapa de filtragem simples, identificamos 1.325 cassetes CRISPR de confiança. Uma análise mais rica foi feita usando o conjunto de dados dos sujeitos diabéticos, com especial atenção às repetições, espaçadores e genes *cas*. O conjunto construído de 362 sequências de repetições únicas foi comparado com CRISPRs conhecidos da base de dados CRISPRdb. Um total de 271 repetições teve *match* na base de dados, e as restantes 91 repetições são potenciais novas repetições para o sistema. As repetições CRISPR foram igualmente submetidos a um agrupamento e classificacção através da ferramenta web CRISPRmap.

Adicionalmente, apenas 37 dos 7,748 espaçadores identificadas do set não redundante, foram acoplados com o seu par genético proveniente de bases de dados de fagos e plasmídeos.

De uma forma geral, a análise computacional da composição CRISPR, em dados de sequenciamento metagenómico, é viável. Fornecendo uma abordagem eficiente para encontrar potenciais novos CRISPRs e para a análise do ecossistema e da história do microbioma humano.

Palavras-chave

CRISPR; Cas; Intestino humano; Microbioma; Metagenoma

Table of Contents

ACKNOWLEDGEMENTS **2**
ABSTRACT **3**
RESUMO **4**
TABLE OF CONTENTS **5**
LIST OF TABLES **8**
LIST OF FIGURES **9**
LIST OF ABBREVIATIONS **11**
INTRODUCTION **12**
LITERATURE OVERVIEW **13**
HUMAN MICROBIOME AND GUT MICROBIOTA 13
NEXT GENERATION SEQUENCING 14
PHAGES 15
Classification *15*
Life Cycles *16*
HOST-PHAGE INTERACTIONS AND BACTERIA DEFENSE MECHANISMS 18
Phage resistance mechanisms *18*
A CLOSER LOOK AT CRISPR-CAS SYSTEMS 20
Brief History of CRISPR research *20*
Structural features of CRISPR-Cas systems *21*
CRISPR-Cas systems as a defense mechanism *26*
Phage adaptation to CRISPR *29*
Current and Future applications *30*
Tools for CRISPR detection and analysis *32*
A CRISPR publicly available database *35*
Tools to further analyse and classify CRISPR arrays *36*
CRISPRs in Metagenomic studies *37*
PROJECT AIM 37
MATERIALS AND METHODS **38**
METAGENOME DATASET 38
IDENTIFICATION AND ANALYSIS OF CRISPR CASSETTES 39
Detection of CRISPR cassettes *39*
Filtering step *39*
REPEAT CLUSTERING AND ANALYSIS 40

Identification of unique repeats and construction of a collection of non-redundant repeat sequences *40*
Comparison to the CRISPR database *40*
CRISPRmap *40*
TAXONOMY OF METAGENOMIC CONTIGS CONTAINING CRISPR CASSETTES 41
CRISPR-associated proteins *41*
IDENTIFICATION OF PROTOSPACERS 42
crAssphage *42*
RESULTS AND DISCUSSION 43
METAGENOMIC DATASET 43
DETECTION AND CHARACTERIZATION OF CRISPR CASSETTES 44
CRISPR Detection software *44*
Filtering step *46*
Predicted CRISPR cassettes *46*
Taxonomy of CRISPR containing contigs *50*
CRISPRmap *54*
Reconstructing a CRISPR array *58*
IDENTIFICATION AND ANALYSIS OF PROTOSPACERS 62
Taxonomy of protospacer origin and compatibility with the CRISPR-cassette taxonomy *62*
Similarity of the spacer composition in the diabetic human gut individual microbiomes *64*
COMPARATIVE ANALYSIS WITH OTHER HUMAN GUT METAGENOMIC DATASETS 65
Repeat Set *65*
Spacers Set *67*
GENERAL CONCLUSION AND FUTURE PERSPECTIVES 68
REFERENCES 69
SUPPLEMENTAL FIGURES 76
SUPPLEMENTAL FIGURE 1 76
SUPPLEMENTAL FILES 77
SUPPLEMENTAL FILE 1 (FILE "SUPPLEMENTAL XCL TABLE 1- CHINESE DNA SAMPLES INFO.XLSX") 77
SUPPLEMENTAL FILE 2 (FILE "SUPPLEMENTAL XCL TABLE 2 – COLLECTION OF PREDICTED RELIABLE CRISPR CASSETTES.XLSX") 77
SUPPLEMENTAL FILE 3: (FILE "SUPPLEMENTAL XCL TABLE 3 – SET OF NON-REDUNDANT REPEATSEQUENCES.XLSX") 77
SUPPLEMENTAL FILE 4: (FILE "SUPPLEMENTAL XCL TABLE 4 – REPEAT SEQUENCE BLAST AGAINST CRISPR DATABASE, CRISPRDB.XLSX") 78
SUPPLEMENTAL FILE 5: (FILE "SUPPLEMENTAL XCL TABLE 5 – SET OF SIGNIFICANT CLUSTERS FROM THE SET OF NON-REDUNDANT REPEATS.XLSX") 78

SUPPLEMENTAL FILE 6: (FILE "SUPPLEMENTAL XCL TABLE 6 – UNIQUE REPEATS COLLECTION AND BLASTX OUTPUT.XLSX") 78

SUPPLEMENTAL FILE 7: (FILE "SUPPLEMENTAL XCL TABLE 7 – CRISPRMAP OUTPUT.XLSX").79

SUPPLEMENTAL FILE 8: (FILE "SUPPLEMENTAL XCL TABLE 8 – COLLECTION OF SPACERS AND CRISPRTARGET OUTPUT.XLSX") 79

SUPPLEMENTAL FILE 9: (FILE "SUPPLEMENTAL XCL TABLE 9 – COMPARATIVE ANALYSIS.XLSX") 79

List of Tables

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. 39

TABLE 2 – CHARACTERISTICS OF IDENTIFIED CRISPR CASSETTES AND SPACERS BY BOTH PILER-CR AND CRT, FOR THE METAGENOMIC DATASET RELATIVE TO THE DIABETIC INDIVIDUALS AND THE HEALTHY INDIVIDUALS..... 47

TABLE 3 - CHARACTERISTICS OF THE AVAILABLE METAGENOMIC DATASETS SUBJECTED TO ANALYSIS FOR CRISPR-CAS CONTENT; * THIS DATASET REFERS TO THE METAHIT PROJECT, AND NOT THE HMP, AS IT IS POSSIBLE TO CHECK FROM THE REFERENCES EXISTENT IN (GOGLEVA, GELFAND AND ARTAMONOVA 2014)..... 65

List of Figures

Figure 1 - Overview of phage classification; C is for circular shape, and L is for linear. Adaptation from (Ackermann 2011). 15

Figure 2 – Schematic representation of a phage life cycle inside a prokaryote; (A) Path is the lytic cycle and path (B) follows the lysogenic cycle. Adaptation from (Botsaris, et al. 2013). 16

Figure 3 - Schematic overview of prokaryotic defence systems. Bacterial cells comprise several independent mechanisms to defend themselves against phage infection (or other invading DNA, such as conjugative plasmids). Defence mechanisms include the blocking of phage adsorption or DNA injection. Other systems act directly on the invader DNA, such as restriction/modification and CRISPR-Cas (CRISPR-associated); Abortive infection systems are usually the ultimate barrier causing cell death upon infection; These defence systems can act independently of each other, or as an ensemble; Note that the infecting blue phage contains a DNA fragment that has a sequence identical to the blue fragment in the CRISPR locus of the host cell. This colouring distinguishes it from the grey phages that are destroyed by other mechanisms. (Westra, et al. 2012). 19

Figure 4 - Timeline overview of the most significant discoveries in CRISPR-Cas research, from 1987 when the repetitive nature of CRISPR arrays was uncovered in E. Coli, until the more recent years where CRISPR-Cas systems were found useful for genome editing; Adaptation from (Louwen, et al. 2014).. 20

Figure 5 – Overview of the schematics of a CRISPR-Cas system; A set of Cas genes is shown in grey, adjacent to a leader sequence (in yellow) that precedes the CRISPR cassette composed of an array of repeats interspaced with spacer sequences, which are DNA sequences of the previous mobile genetic invaders (phages, plasmids). 21

Figure 6 - Example of a palindrome in a genomic sequence; The DNA locus shows a sequence which is identical on each 5'-to-3' DNA strand. The sequence is the same when one strand is read left to right and the other strand is read right to left. Recognition sites of many restriction enzymes are palindromic. 22

Figure 7 – Schematic overview of Cas proteins classification modules according to their function; Protein names follow the current nomenclature; Dispensable components are indicated by dashed outlines. Cas6 is shown with a solid outline for type I because it is dispensable in some but not most systems and by a dashed line for type III because most systems lack this gene and use the Cas6 provided in trans by other CRISPR–cas loci; The two colours for Cas4 and three colours for Cas9 reflect that these proteins contribute to different stages of the CRISPR–Cas immunity; The functions shown for type IV and type V system components are proposed based on homology to the cognate components of other systems, and have not yet been experimentally verified. (Makarova, Wolf and Alkhnbashi, et al. 2015) 24

Figure 8 - Schematics of the CRISPR-Cas system immune system; (Stage 1), CRISPR RNA biogenesis (Stage 2) and target interference (Stage 3); 26

Figure 9 – Example of an output from CRT showing a predicted CRISPR cassette between two different contigs, evidencing the problem with CRT regarding the non individualization of the contigs. 44

Figure 10 - The case where it is possible to observe that both software's output provides a 100% identical repeat sequence but the number of spacers predicted for the cassette differentiates in more than two units; Output from CRT (on the top) detects an array of 10 repeats and 9 spacers, while the output from PILER-CR (on the bottom) detects 8 repeats and 7 spacer sequences. 48

FIGURE 11 - 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. 51

FIGURE 12 – REPEAT ALIGNMENT WITH WEBPRANK (HTTP://WWW.EBI.AC.UK/GOLDMAN-SRV/WEBPRANK/); THE COLOUR CODE IS USED TO DISTINGUISH BETWEEN NUCLEOTIDES; RED IS FOR T; BLUE IS FOR A; GREEN IS FOR C AND YELLOW IS FOR G. 52

FIGURE 13 – DETAILS ON FAMILY 5 PROVIDED BY THE OUTPUT FROM CRISPRMAP; WEBLOGO TOOL IS USED TO GENERATE SEQUENCE LOGOS OF ALL ALIGNED FAMILY MEMBERS. SEQUENCE LOGOS PRESENT A GRAPHICAL REPRESENTATION OF A MULTIPLE SEQUENCE ALIGNMENT OF FAMILY 5 MEMBER REPEAT SEQUENCES, WHERE NUCLEOTIDES LETTERS THAT STAND OUT REPRESENT THE MOST CONSERVED POSITIONS. 55

FIGURE 14 – DETAILS ON MOTIF 3 PROVIDED BY THE OUTPUT FROM CRISPRMAP; NOTE THAT THE MOTIF WAS UPDATED AFTER THE INPUT OF OUR NEW SET OF SEQUENCES, NOT PRESENT IN THEIR DATABASE (48 NEW REPEAT SEQUENCES); IT IS POSSIBLE TO OBSERVE THE PALINDROMIC CHARACTERISTIC OF CRISPR REPEATS IN THE HAIRPIN-LIKE STRUCTURE OF THE MOTIF 55

FIGURE 15 - 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. 59

FIGURE 16 - SCHEMATICS OF THE ARCHITECTURE OF AN IDENTIFIED TYPE II CRISPR-CAS SYSTEM, FROM CONTIG 'SCAFFOLD7464_7, ORIGINATED IN INDIVIDUAL DOM016; CONTIG HAS A TOTAL SIZE OF 10,085 BP; (A) THE ANNOTATED CAS GENES, DOWNSTREAM OF THE ARRAY, ARE CONSTITUTED BY *CAS9* GENE, CHARACTERISTIC OF TYPE II SYSTEMS, FOLLOWED BY THE CORE CAS GENES, CAS1, CAS2 AND CAS4; (B) THE CRISPR ARRAY, WITH LENGTH 322BP, IS COMPOSED OF 4 SPACERS (BLUE COLOURED RECTANGLES) AND 5 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 'OLDEST' ONE, AND SPACER 4 IS THE NEWEST (IT WAS THE LAST ONE TO BE INSERTED INTO THE ARRAY); NOTE THAT THE SCHEME IS MADE ON SCALE AND DOES NOT REPRESENT THE EXACT SIZE OF THE GENES OR CRISPR ARRAY. 60

FIGURE 17 - SCHEMATICS OF THE ARCHITECTURE OF AN IDENTIFIED TYPE V PUTATIVE CRISPR-CAS SYSTEM, FROM CONTIG 'SCAFFOLD23627_1', ORIGINATED IN INDIVIDUAL DLM001; CONTIG HAS A TOTAL SIZE OF 6502 BP; (A) THE ANNOTATED CAS GENES, UPSTREAM OF THE ARRAY, ARE CONSTITUTED BY *CPF1* GENE, CHARACTERISTIC OF TYPE V SYSTEMS, FOLLOWED BY THE CORE CAS GENES, CAS4, CAS1 AND CAS2; (B) THE CRISPR ARRAY, WITH LENGTH 2161BP, IS COMPOSED OF 29 SPACERS (COLOURED RECTANGLES) AND 30 REPEAT SEQUENCES (◆), WITH AN ADJACENT LEADER SEQUENCE UPSTREAM OF THE FIRST REPEAT; SPACER 1 IS THE ONE FURTHEST FROM THE LEADER SEQUENCE SO IT IS THE 'OLDEST' ONE, AND SPACER 29 IS THE NEWEST; NOTE THAT THE SCHEME IS MADE ON SCALE AND DOES NOT REPRESENT THE EXACT SIZE OF THE GENES OR CRISPR ARRAY 61

FIGURE 18 - THE OUTPUT FROM BLASTN ALGORITHM AFTER 'BLASTING' OUR ENTIRE SET OF NON-REDUNDANT SPACERS AGAINST THE 97 KBP GENOME OF BACTERIOPHAGE CRASSPHAGE; 'DOM013_292' IDENTIFIES THE SPACER HAS BEING FROM INDIVIDUAL SAMPLE DOM13 63

List of abbreviations

Abi	Abortive Infection
A	Adenyne
iap	alkaline phosphatase isozyme conversion protein
AFLP	Amplified Length Polymorphisms
bp	base-pairs
BMI	Body Mass Index
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	complementary RNA
Cas	CRISPR associated proteins
CRISPRdb	CRISPR database
C	Cytosine
DNAase	Deoxyribonuclease
DNA	Deoxyribonucleic Acid
DR	Direct Repeats
DDBJ	DNA Databank of Japan
dsDNA	double stranded DNA
GOS	Global Ocean Sampling
G	Guanyne
GM	Gut Microbiome
IBD	Inflammatory Bowel Disease
RNAi	interference ribonucleic acid
int-crRNA	intermediate complementary RNA
mat-crRNA	mature complementary RNA
MFE	Minimum Free Energy
MGE	Mobile Genetic Element
MLST	Multilocus Sequence Typing
MSA	Multiple Sequence Alignment
NGS	Next-Generation Sequencing
ncRNA	non-coding RNA
PCA	Principal component analysis
PDB	Protein Databank
PAM	Protospacer-adjacent motif
RM	Restriction Modification
RNAase	Ribonuclease
RNA	Ribonucleic Acid
rRNA	ribosomal Ribonucleic Acid
spp.	species
Sie	Superinfection exclusion
3D	Three-Dimensional
tracrRNA	transactivating CRISPR RNA
T	Tymine
U	Uracile
WGS	Whole-Genome Shotgun

Introduction

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 microbiomes. Most efforts in metagenome study 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.

Literature Overview

Human microbiome and Gut Microbiota

The Human microbiome, a complex community of symbiotic, pathogenic and commensal miccrooganisms (Lederberg 2000), represents the collection of all microbes that live on and within a human being. At 10^{14} species, comprising at least 20 million unique microbial genes, the microbiome constitutes the largest genetic component of the human super organism (Hoffman, et al. 2015).

The majority of the bacteria resides in the gastrointestinal tract, or gut, and have a profound inffluence on human physiology and nutrition, and are crucial for human life (J. Qin, R. Li, et al. 2010) (Guinane and Cotter 2013). The human gut microbiome (GM), composed of 10 times more cells than human cells, making up to 100 trillion cells (Guinane and Cotter 2013), might even be considered as an organ within an organ (Eckburg, et al. 2005). The human host provides a nutrient-rich environment and the microbiota provides indispensable functions that humans cannot exert themselves, such as the production of some vitamins (LeBlanc, et al. 2013), digestion of complex polysaccharides (Cantarel, Lombard and Henrissat 2012) and the shaping of the immunological environment (Round and Mazmanian 2009) (Kelly and Mulder 2012).

The GM has the potential for multistability and coevolves with us (Quercia, et al. 2014). Changes to this population can have major consequences, both beneficial and harmful, for human health. Indeed, it has been suggested that disturbance of the gut microbiota can be significant with respect to pathological intestinal conditions such as obesity and malnutrition, systematic diseases such as diabetes and chronic inflammatory diseases such as inflammatory bowel disease (IBD), which can be classified into ulcerative colitis (UC) and Crohn's disease (CD).

Some recent analysis demonstrate that the microbiota is relatively stable over time, but describes an evolutionary trajectory along the course of human life. In fact, the GM ecossystem changes its structural and functional layout from early infancy to old age (Yatsunenko, et al. 2012) (Salazar, et al. 2014), however it is very specific for each individual. This specificity is due to environmental pressure and some endogenous factors, such as diet, age, host genetics, and physiological states (Candela, et al. 2012).

Next Generation Sequencing

Traditional microbiology, despite its achievements, is limited, and mostly based on growing pure cultures of species in specific media, i. e., heavily culture-based. Even other approaches, making use of the 16S rRNA gene sequencing, although immensely important in the beginning phases of microbiome study and in helping scientists define evolutionary relationships among bacteria, also carry limitations (Hollister, Brooks and Gentry 2015). The reality is that a huge, undeniable part, of microorganisms doesn't have cultivated representatives in several environments (Rappé and Giovannoni 2003). In fact, approximately 20 to 60% of all microorganisms that inhabit the human body are considered to be uncultivable (The NIH HMP Working Group, et al. 2009).

Metagenomics, defined as the functional and sequence-based analysis of the collective microbial genomes that are contained in an environmental sample (Metagenomics: Sequences from the Environment 2006) appeared as the answer to the existing limitations and was initially applied in several studies of environmental microbial communities, including aquatic and soil ecosystems (Tyson , et al. 2004) (Rusch 2007).

The introduction of culture-independent approaches allowed to increase the understanding of the complexity and diversity of the microbiota, but it was the revolution of Next-Generation Sequencing (NGS) technologies (Metzker 2010), such as 454 pyrosequencing, SOLiD™ (sequencing by ligation), and Illumina (sequencing by synthesis) that intensely accelerated the development of metagenomic projects, thanks to their reduced costs and larger processing speed and power. Alongside the development of bioinformatic analytical tools, it was allowed researchers a better understanding of the microbial communities of interest from a phylogenetic and functional perspective.

As a result of these advances in sequencing technologies and the ever-growing interest in the study of the human microbiome, two large-scale sequencing projects were initiated: the Human Microbiome Project (HMP) and the Metagenomics of the Human Intestinal Tract (MetaHIT) project. The first, focus on the diversity of different body site's microbiomes, including the gastrointestinal tract, in the healthy human population (NIH HMP Working Group et al. 2009), while MetaHIT focus on the correlation between the variation of the gut microbiome (with population, genotype, disease, age, nutrition, medication, and environment) and intestinal diseases, such as obesity and IBD (J. Qin, R. Li, et al. 2010).

Phages

Bacteriophages, or phages, are the most abundant and diverse organisms in the biosphere and they are a ever-present feature of prokaryotic life. They are viruses capable of infecting a specific target bacteria and, in some cases, lyse bacterial cells.

Classification

Bacteriophages come in different sizes and shapes but most of them have the same basic features: a head or capsid and a tail. A bacteriophage's head structure, regardless of its size or shape, is made up of one or more proteins which coats the nucleic acid. Although there are some phages that don't have a tail, most of them do have one attached to its head structure. The tail, a hollow tube, is an important structural feature during infection since it allows the passage of the phage nucleic acid (Krupovic, et al. 2011).

Virus can be double or single-stranded DNA or RNA, can be tailed or polyhedral and finally, filamentous or pleomorphic. Accordingly, based on the type of nucleic acid and virion morphology, bacterial viruses are classified by the ICTV (International Committee on Taxonomy of Viruses). The latest classification included 11 different order and families (See Figure 1).

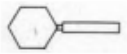
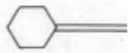
Shape	Order or family	Nucleic acid, particulars, size	Member
	Caudovirales	dsDNA (L), no envelope	
	*Myoviridae*	Tail contractile	T4
	*Siphoviridae*	Tail long, noncontractile	λ
	*Podoviridae*	Tail short	T7
	*Microviridae*	ssDNA (C), 27 nm, 12 knoblike capsomers	ϕX174
	*Corticoviridae*	dsDNA (C), complex capsid, lipids, 63 nm	PM2
	*Tectiviridae*	dsDNA (L), inner lipid vesicle, pseudo-tail, 60 nm	PRD1
	*Leviviridae*	ssRNA (L), 23 nm, like poliovirus	MS2
	*Cystoviridae*	dsRNA (L), segmented, lipidic envelope, 70–80 nm	ϕ6
	*Inoviridae*	ssDNA (C), filaments or rods, 85–1950 x 7 nm	fd
	*Plasmaviridae*	dsDNA (C), lipidic envelope, no capsid, 80 nm	MVL2

Figure 1 - Overview of phage classification; C is for circular shape, and L is for linear. Adaptation from (Ackermann 2011).

Life Cycles

Outside their hosts, viruses are inert objects and cannot replicate. For this reason, phages depend on the host to propagate. Once in contact with their host, the bacteriophage attaches itself to the bacteria's cell wall, through a host-specific receptor found on the bacteria's surface (Labrie, Samson and Moineau 2010) .

Completing the adsorption step, the phage injects its genetic material (nucleic acid) into the host cell. From here, phages have two possible life cycles, which will dictate their role in bacterial biology. Lytic and lysogenic cycles are the two different methods of viral replication.

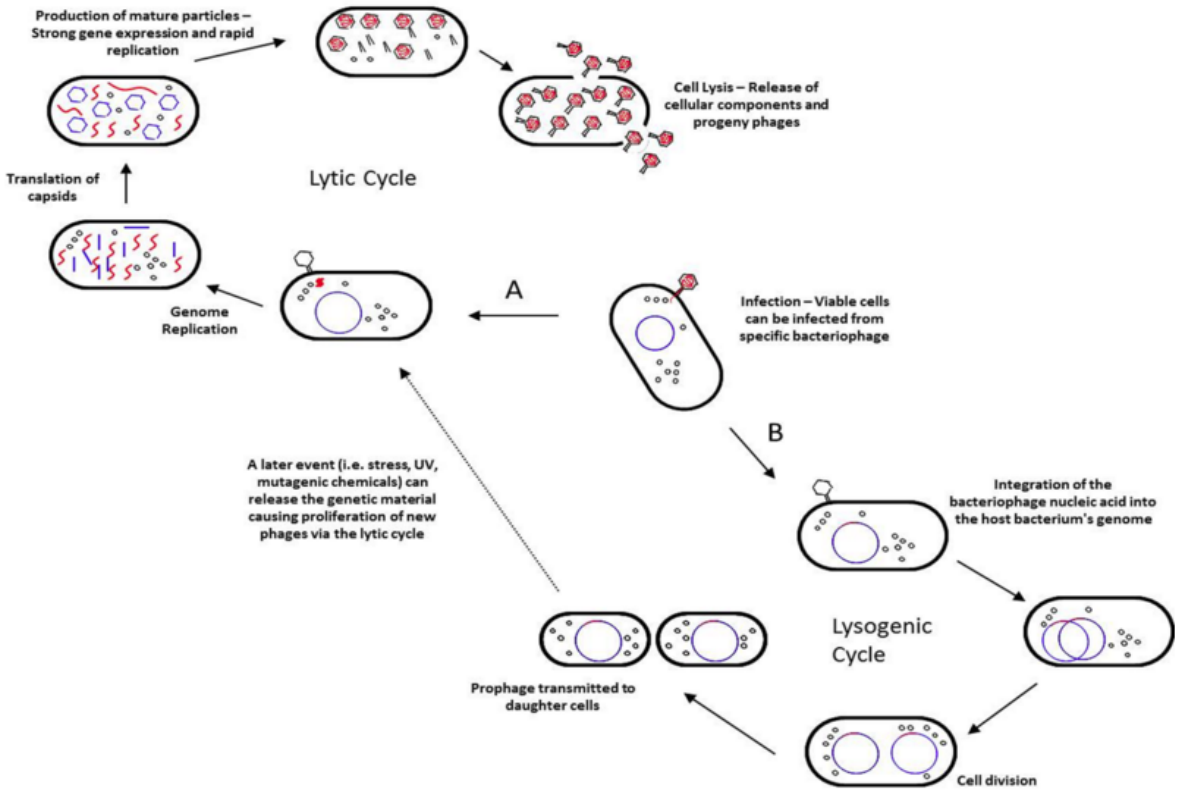


Figure 2 – Schematic representation of a phage life cycle inside a prokaryote; (A) Path is the lytic cycle and path (B) follows the lysogenic cycle. Adaptation from (Botsaris, et al. 2013).

Lytic Cycle

During a lytic cycle, or “virulent phages” cycle, phages infect and rapidly kill their host cells. When the infection is successful, the phage will make use of the host cell’s chemical energy as well as its biosynthetic machinery to produce phage nucleic acids and phage proteins. The bacterial metabolism is shut down, the viral genome is transcribed and assembled into virons. The phages are capable of releasing enzymes that attack and destroy the bacterial membrane, such as lysozymes, endopeptidases or amidases (Clokie, et al. 2011) (Samson, et al. 2013). When cell lysis occurs, newly assembled phages are released and proceed to infect other neighbour bacterial cells. The cycle continues until all target cells have been lysed.

Lysogenic cycle

The lysogenic life cycle in contrast, is where phages, instead of directly killing their hosts, assume prophage form, integrating themselves into the host genome, or existing as plasmids within their host cell (Clokie, et al. 2011). When the bacterium reproduces, the prophage is replicated along with the host chromosomes, coexisting as opposed to lysing the host cell, in a cycle that can be stable for several generations. A regulator gene produces a repressor protein that suppresses the lytic activity of the phage, but various environmental factors, such as ultraviolet irradiation may prevent synthesis of the repressor, leading to normal phage development and lysis of the bacterium.

Called ‘lysogeny’, this phenomenon may even alter the host phenotype and fitness: by providing immunity against infection by further phage particles of the same type, ensuring that there is only one copy of phage per bacterial cell; acting as anchor points for genome rearrangements via gene disruption and introduction of new fitness factors (by transduction[1] and/or lysogenic conversion[2]). In fact, this phage integration process may also result in the acquirement, by the host bacterial cell, of genes that can confer pathogenicity or increase virulence. (Jassim and Limoges 2014)

Lysogenic phages, or temperate phages, act, indeed, as a powerful driving force in bacterial diversity, adaptation and evolution.

After numerous cell divisions, one cell can spontaneously lyse and release progeny phages.

[1] The process, by which, DNA is transferred from one bacterium to another, with the aid of a virus.

[2] The process by which the phage induces changes in the host bacteria, at a phenotypic level, imparting genes with functions not present before.

Host-phage interactions and bacteria defense mechanisms

Phages are arguably the most abundant biological entity on the planet, even outnumbering bacterial cells by a factor of 10:1 in several natural ecosystems (Chibani-Chennoufi, et al. 2004). Their various distributions and abundance have a significant impact on microbial ecology and evolution. In fact, coevolution between phages and bacteria has been shown to have an important role in maintaining phenotypic and genotypic diversity, increase the rate of bacterial and phage evolution and divergence, affect the community structure, and shape the development of bacterial traits (Koskella and Brockhurst 2014).

Being responsible for a significant mortality rate in their hosts (Clokie, et al. 2011), predation by phages presents itself as an important challenge to bacterial survival, and bacteria have evolved numerous mechanisms to resist phage infection.

Phage resistance mechanisms

Phage resistance mechanisms in bacteria can be passive or active. There are three important anti-viral systems responsible for: prevention of phage adsorption, preventing phage DNA entry and cutting phage nucleic acids.

The phage infection process begins with the specific adsorption of the phage to a receptor on the host surface. As a first line of defence, bacteria can disrupt this phage adsorption by eliminating or masking the corresponding receptors. This can be achieved in different ways: by the formation of a extracellular matrix, acting as a physical barrier, between phages and their receptors; by the production of competitive inhibitors that will bind to these receptors making them unavailable for phages; or restructuring of the cell surface receptors or its 3D conformation. In some cases, a second layer of defence extends to the blockage of the injection of phage DNA, through the use of membrane associated proteins belonging to a Superinfection exclusion (Sie) system, or by reinforcing the cell wall, thickening the peptidoglycans (Hyman and Abedon 2010) (Labrie, Samson and Moineau 2010).

Following successful adsorption and penetration of the phage particle, bacteria will resort to restriction-modification (RM) systems or CRISPR-Cas systems.

RM systems consist of two or more genes encoding a sequence-specific restriction endonuclease that cleaves invading unmethylated DNA, and a DNA methylase that modifies specific bases of the bacterial genetic material to protect it from the endonuclease action. When an unmethylated phage DNA enters a cell, it will be either recognized by the restriction enzyme and “cut” in specific sites or, methylated to avoid restriction, hence leading to the initiation of the phage’s lytic cycle. In the latter, new virons turn out to be insensitive to the equivalent restriction enzyme and can infect new bacteria harbouring the same RM system The phage will remain insensitive until it infects a bacteria that does not encode the same methylase gene, in which case the new virons will become unmethylated again and become susceptible to the RM system of the original bacteria (Hyman and Abedon 2010).

The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and Cas (CRISPR-associated proteins) system is the only known bacterial adaptive and heritable immune system (van

der Oost, et al. 2009). Short phage sequences are captured and incorporated in CRISPR arrays, which can be transcribed and processed into small RNAs. The latter, with the aid of a multifunctional protein complex (Cas proteins), are used to recognize and destroy the recurring invader (Barrangou, Fremaux, et al. 2007). Conceptually, many aspects of the CRISPR-Cas system are similar to adaptive mechanisms of small RNA-based defence that protect eukaryotic cells.

Finally, abortive infection systems interfere with various aspects of phage replication and packaging, while leading to death of the host. Abortive infection (Abi) systems are the ultimate resistance mechanism that can lead to the death of the host cell. Abi systems target the late stages of phage development, which include replication, transcription, translation and phage packaging (Labrie, Samson and Moineau 2010).

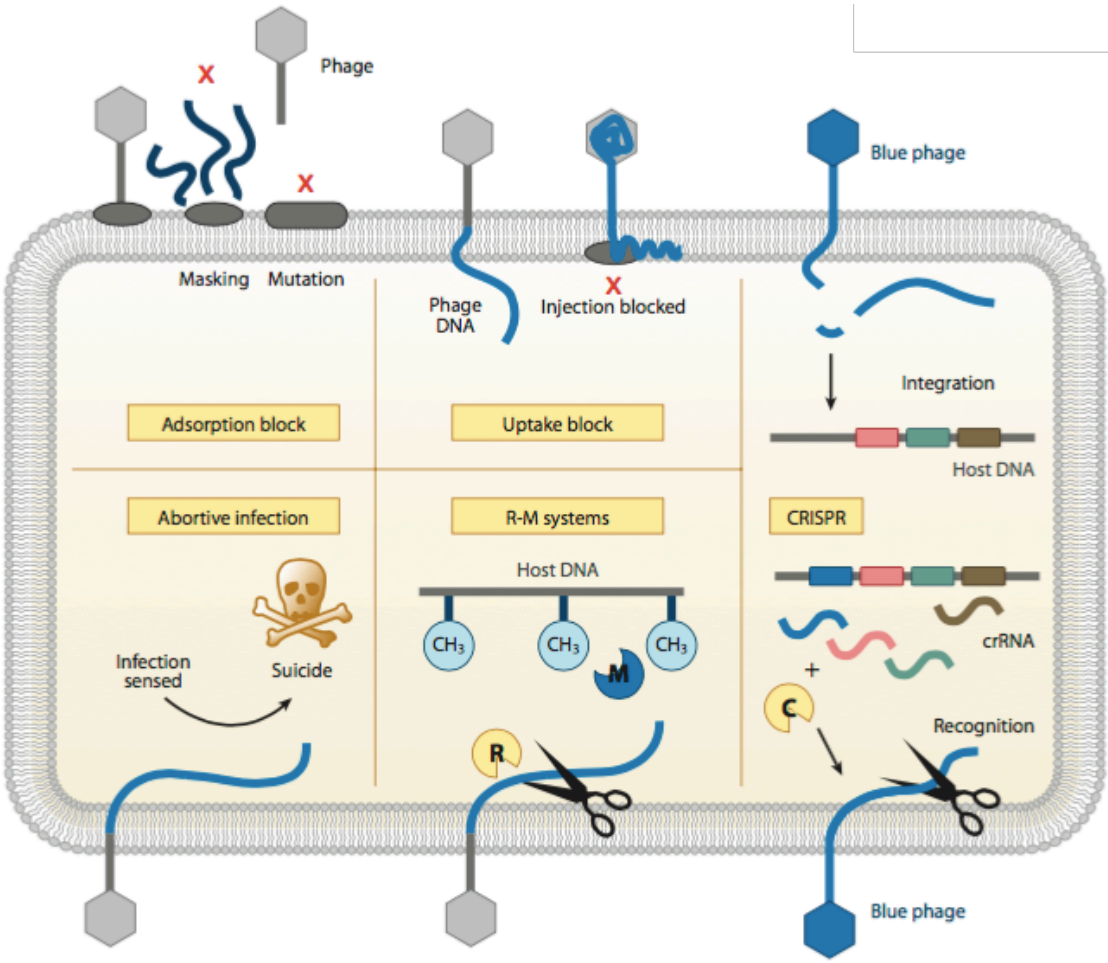


Figure 3 - Schematic overview of prokaryotic defence systems. Bacterial cells comprise several independent mechanisms to defend themselves against phage infection (or other invading DNA, such as conjugative plasmids). Defence mechanisms include the blocking of phage adsorption or DNA injection. Other systems act directly on the invader DNA, such as restriction/modification and CRISPR-Cas (CRISPR-associated); Abortive infection systems are usually the ultimate barrier causing cell death upon infection; These defence systems can act independently of each other, or as an ensemble; Note that the infecting blue phage contains a DNA fragment that has a sequence identical to the blue fragment in the CRISPR locus of the host cell. This colouring distinguishes it from the grey phages that are destroyed by other mechanisms. (Westra, et al. 2012).

A closer look at CRISPR-Cas Systems

Brief History of CRISPR research

The first report that described clustered repeats, in 1987, found an array of 14 similarly sized repeats interspersed by non-repeating sequences, adjacent to an *iap* gene in *Escherichia coli* K-12 chromosome (Ishino, et al. 1987), but at the time the physiological role of the repetitive DNA remained unknown. In subsequent years similar arrays were found in other bacterial species like *Mycobacterium tuberculosis* (Hermans, et al. 1991) and Archaea species like *Haloferax mediterranei* (Mojica, Ferrer, et al. 1995). The accumulation of sequenced microbial and Archaea genomes allowed genome-wide computational searches for clustered repeat arrays, being the ever first analysis done by Mojica *et al* in 2000 (Mojica, Diez-Villasenor and Soria, et al. 2000). At this time, this type of repeat structure was termed Short Regularly Spaced Repeats (SRSR). SRSR were only renamed CRISPR in 2002 upon discovery of CRISPR-associated genes (Cas genes) that were found almost adjacent to the repeat arrays (Jansen, et al. 2002). These genes encoded nuclease or helicase proteins, which are enzymes that can cut DNA. In subsequent studies it was uncovered 24-45 additional Cas genes appearing in close proximity to the arrays (Haft, et al. 2005) (Makarova, Grishin, et al. 2006).

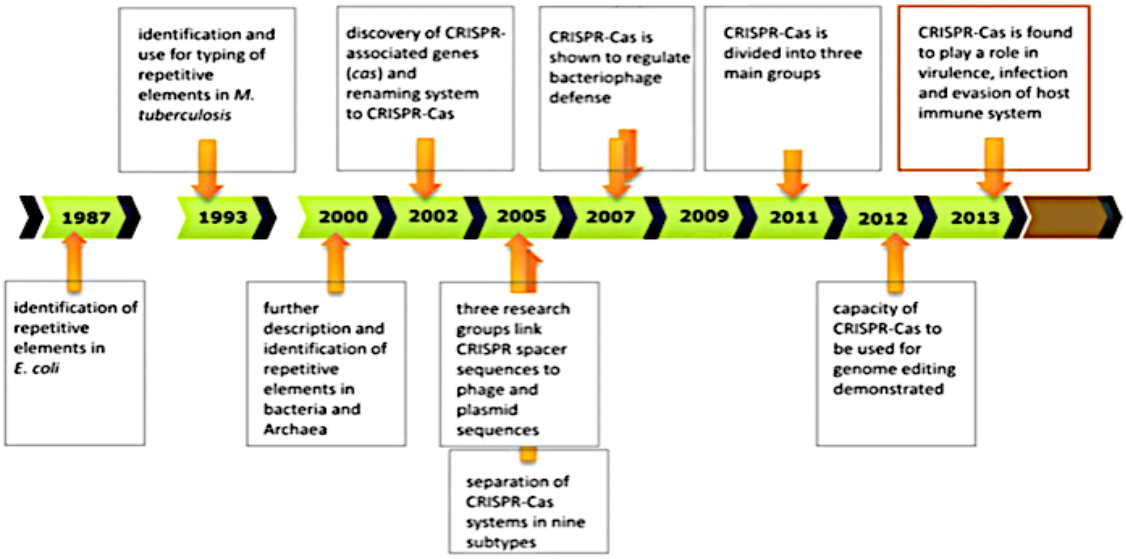


Figure 4 - Timeline overview of the most significant discoveries in CRISPR-Cas research, from 1987 when the repetitive nature of CRISPR arrays was uncovered in E. Coli, until the more recent years where CRISPR-Cas systems were found useful for genome editing; Adaptation from (Louwen, et al. 2014).

Only in 2005 was attributed significance to the non-repeating regions of the CRISPR arrays, when it was found that they derived from phage DNA and extra chromosomal DNA such as viruses and plasmids (Pourcel, Salvignol and Vergnaud 2005) (Mojica, Diez-Villasenor and Garcia-Martinez, et al. 2005) (Bolotin, et al. 2005). From this discoveries it was posteriorly hypothesized a role of CRISPR-Cas in mediating an immunological response against foreign elements, using spacers as templates for RNA molecules, analogous to the RNAi system used by eukaryotic cells (Barrangou, Fremaux, et al. 2007).

More recently, around 2012, the CRISPR huge potential in genome engineering, working as an editing tool in human cell cultures, was unravelled (Jinek , et al. 2012).

Structural features of CRISPR-Cas systems

Clustered regularly interspaced short palindromic repeats (CRISPRs) are defined as an array of short direct repeats (21-50 bp) interposed with spacer sequences (20-72 bp). Together with Cas genes, they form the CRISPR-Cas system, varying greatly among microbial and archaea species (Kunin, Sorek and Hugenholtz 2007). The CRISPR-Cas system works as a prokaryotic immune system that confers resistance to alien genetic elements such as phages and plasmids, providing a form of acquired immunity. In the first stage of bacterial defence, bacteria incorporate fragments of the invasive mobile elements as novel spacers. These spacers are then transcribed into small RNAs, which, together with the Cas protein complex, guide the way to interfere with phage replication and then efficiently remove the potential treat (Horvath and Barrangou 2010).

CRISPRs were found in approximately 50% of sequenced bacteria genomes and 95% of sequenced Archaea (Grissa, Vergnaud and Pourcel 2007) (Jore, Brouns and van der Oost 2012).

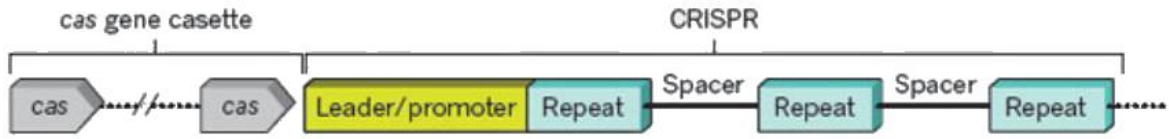


Figure 5 – Overview of the schematics of a CRISPR-Cas system**;** A set of Cas genes is shown in grey, adjacent to a leader sequence (in yellow) that precedes the CRISPR cassette composed of an array of repeats interspaced with spacer sequences, which are DNA sequences of the previous mobile genetic invaders (phages, plasmids).

Leader

A non-coding sequence of approximately 100 - 500 bp is located 5' to most CRISPR arrays, directly adjoining the first repeat (Wei, et al. 2015). This, usually, adenine/thymine (A/T)-rich sequence (Jansen, et al. 2002) (Erdmann and Garrett 2012), is suggested to contain transcriptional promoters for the transcription of CRISPR loci into small RNA (Makarova, Grishin, et al. 2006). Much like the repeats themselves, leaders can be identical within a genome, especially if several CRISPR arrays are found within the same chromosome (Jansen, et al. 2002), but can be quite dissimilar among species. When a new repeat-spacer unit is added to the CRISPR array, it is typically integrated between the leader and its previous adjacent unit, suggesting that the leader might also function as a recognition sequence for the addition of new spacers (Barrangou, Fremaux, et al. 2007).

Repeats

In a single array, repeats, with an average length of 32 bp, are usually identical, sequence and size wise (Makarova, Haft, et al. 2011). Related species can have similar repeat sequences, but the

overall sequence diversity of repeats in bacteria and archaea is great (Kunin, Sorek and Hugenholtz 2007), although it has been shown that CRISPR arrays with the same repeat and Cas gene set can be found in multiple bacterial species, implying horizontal gene transfer (HGT) (Godde and Bickerton 2006). Despite their sequence diversity, a majority of CRISPR repeats have the potential to form stable RNA secondary stem-loop-like structures, meaning they are partly palindromic, containing a short, 5 – 7 bp palindrome[3] (hence the word "palindromic" in the CRISPR acronym) (Kunin, Sorek and Hugenholtz 2007).

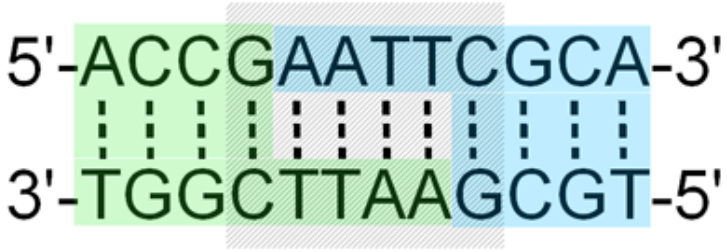


Figure 6 - Example of a palindrome in a genomic sequence; The DNA locus shows a sequence which is identical on each 5'-to-3' DNA strand. The sequence is the same when one strand is read left to right and the other strand is read right to left. Recognition sites of many restriction enzymes are palindromic.

Kunin and colleagues (Kunin, Sorek and Hugenholtz 2007) observed that substitutions in the predicted stem structure are consistently accompanied by compensatory changes that preserve the base pairing, and together with conservation of G-U pairs in the stem, they implied that the structures can form in the transcriptional products of the locus, highlighting the importance of this hairpin structure for the functionality of CRISPRs. Apart from this structural feature, many repeats appear to have a conserved 3' terminus of GAAA(C/G), suggested to act as a binding site for one or more CRISPR-associated proteins (Godde and Bickerton 2006) (Kunin, Sorek and Hugenholtz 2007).

The number of repeat-spacer units that constitute a CRISPR array is also variable, ranging from as few as two to as many as several hundred, being the largest CRISPR found so far, observed in *Haliangium ochraceum* DSM 14365 with 588 repeats. The largest repeat to date has a length of 50 bp and was identified in *Weeksella virosa* (1 CRISPR array harbouring 20 spacers) (Pourcel and Drevet 2013).

Spacers

In any CRISPR array, spacers are typically unique, with an average size similar to the size of the repeats belonging to that same array (around 36 bp) (Pougach, Lopatina and Severinov 2012). One

[3] A palindrome is a nucleic acid sequence (DNA or RNA) that is the same whether read 5' (five-prime) to 3' (three-prime) on one strand or 5' to 3' on the complementary strand with which it forms a double helix.

of the focus points in all the CRISPR-Cas system is the spacer's sequence that has its origins in phage's genomes and other extrachromossomal elements. The adaptation against invading viruses depends on the insertion of these foreign nucleic acid-derived sequences into the CRISPR array as novel spacers, and it's the transcript of the spacer that guides a cleavage protein complex to cut out the DNA of the invader.

The selection of spacer precursors (protospacers) (Deveau, et al. 2008) from the invading DNA seems to be determined, for some bacteria, by the recognition of protospacer-adjacent motifs (PAMs). In some types of CRISPR–Cas system, cleavage only occurs if the protospacer is flanked by this motif (Heler, et al. 2015). These motifs are usually only 2 – 5 bp long, starting immediately or up to four positions of one extremity of the protospacer. They differ between variants of the CRISPR-based system and organism. The existence of this recognizing motif, together with the polarized arrangement of the spacers in the CRISPR arrays, and the conservation of their relative orientation with respect to the PAM sequences, have fundamental implications for the spacer uptake process (Mojica, Díez-Villaseñor, et al. 2009).

While CRISPR-Cas systems primarily add novel spacers to the array, internal spacers sometimes are deleted. These deletions help to limit the size of CRISPR array, maintaining relatively small numbers of repeat-spacer units, conserving a certain genomic homeostasis and helping to limit transcript lengths. Deleting spacers from phages that were encountered in the past, located near the 3' end, would allow conservation of spacers targeting viruses currently circulating. Moreover, proximal spacers (closer to the 5' end) are probably transcribed at a higher frequency than distal spacers (closer to 3' end).

Multiple spacer acquisition from the same origin is less frequent, as are internal insertions (Makarova, Haft, et al. 2011).

In some bacteria, as a form of autoimmunity, CRISPR arrays may harbour self-targeting spacers (Stern, Keren, et al. 2010). These spacers are widely distributed over diverse phylogenetic lineages, and are dispersed throughout different arrays in each organism, occurring in approximately 18% of all CRISPR-bearing organisms. It is suggested that the lack of conservation of these spacers across species, together with the co-occurrence of degraded repeats near self-spacers, indicate that the acquisition of this type of spacers is damaging for the stability and integrity of the host organism. In fact, the self-targeting spacer, the targeted gene, or even the entire CRISPR-Cas is prone to be lost.

The existence of this type of spacers may explain why such a valuable bacterial system is only present in approximately half of bacteria.

However, it cannot be ruled other potential role for these self-targeting spacers in which they act in a type of transient regulation. For instance, an alternative model is that CRISPR targets endogenous host genes that contribute to virus replication (R. Barrangou 2015).

Cas proteins

CRISPR-associated proteins (Cas proteins) play a significant role in the CRISPR-Cas systems immunity mechanism.

The core defining feature of CRISPR-Cas types and subtypes are the cas genes and the proteins they encode, which are highly diverse in function and structure, illustrating the many biochemical functions that they carry throughout the different steps of CRISPR-activity. CRISPR-Cas systems have been classified into three major types, namely type I, type II and type III, and two putative new types, type IV and V. These five types are divided into 16 subtypes, given their genetic content and structural and functional differences (Makarova, Wolf and Alkhnbashi, et al. 2015).

Genetically, Cas1 and Cas2 universally occur across types and subtypes, neverthless, the three main types are readily distinguishable by virtue of the presence of unique signature proteins: Cas3 for type I, Cas9 for type II and Cas10 for type III (Makarova, Haft, et al. 2011). See **Supplementary Figure 1** for a complete overview on all the CRISPR-Cas types and subtypes according to the most recent classification.

The Cas proteins can also be divided into four distinct functional modules relative to the different stages in CRISPR immunization activity: adaptation (for spacer acquisition); expression (crRNA processing and target binding); interference (target cleavage); and ancilliary (regulatory and other CRISPR-associated functions).

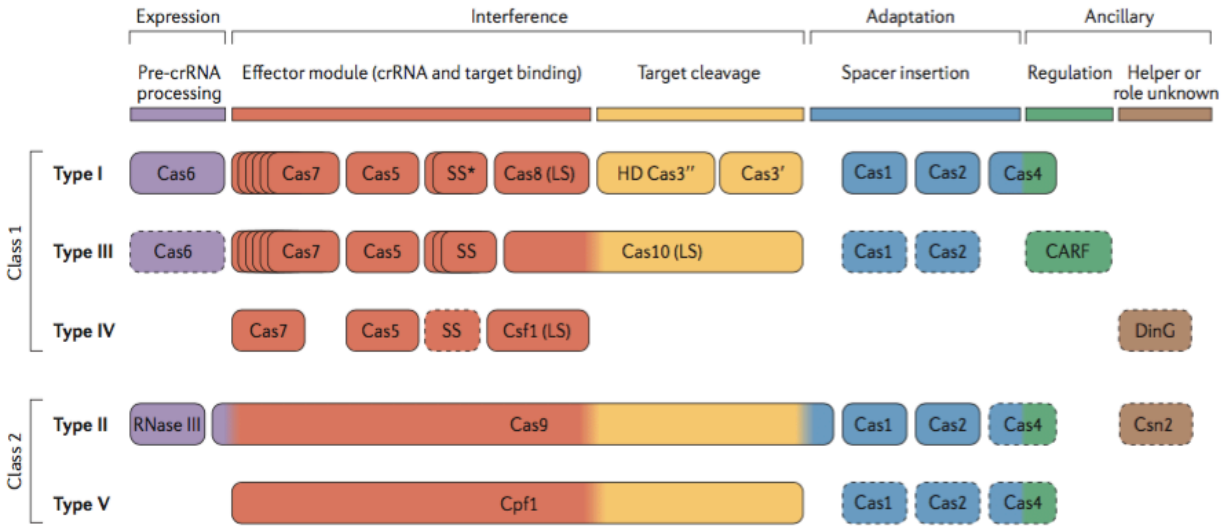


Figure 7 – Schematic overview of Cas proteins classification modules according to their function; Protein names follow the current nomenclature; Dispensable components are indicated by dashed outlines. Cas6 is shown with a solid outline for type I because it is dispensable in some but not most systems and by a dashed line for type III because most systems lack this gene and use the Cas6 provided in trans by other CRISPR–cas loci; The two colours for Cas4 and three colours for Cas9 reflect that these proteins contribute to different stages of the CRISPR–Cas immunity; The functions shown for type IV and type V system components are proposed based on homology to the cognate components of other systems, and have not yet been experimentally verified. (Makarova, Wolf and Alkhnbashi, et al. 2015)

The adaptation module is largely uniform across CRISPR-Cas systems and consists of the Cas1 and Cas2 proteins, with possible involvement of protein Cas4 (restriction endonuclease), for all types except type II, where it is constituted by Cas9 protein. Cas1 is an integrase responsible for mediating the insertion of new spacers into the CRISPR array by cleaving specific sites within the repeats. The role of Cas2 isn't as well understood, although it has been shown to have DNAase and RNAase activities (enzymes that catalyse the degradation of nucleic acids). This homologue of the mRNA interferase toxins, always appears associated with the Cas1 integrase, and its presence has been found to be a requirement in the *Escherichia coli* type I system, where it forms a complex with Cas1 (Nunez, et al. 2014).

The expression and interference modules are represented by multisubunit CRISPR RNA (crRNA)-effector complexes, or, in type II systems, by a single large protein Cas9. During the expression phase, in a step catalysed by a RNA endonuclease, the pre-crRNA is bound to the multisubunit complex, or to Cas9, and processed into a mature crRNA. Usually, in type I and III systems, the catalyst is the Cas6 protein, and in type II systems the same role is completed by an alternate mechanism involving an RNAase and a transactivating CRISPR RNA (tracrRNA).

During the interference stage, nuclease activity comes into play. Depending on the CRISPR type, the cleavage of the target sequence (protospacer) is catalysed by different proteins: the HD family nuclease in type I systems (Cas 3 associated); in type III systems, by the action of Cas7 and Cas10; and by Cas9 in type II systems (Makarova, Wolf and Alkhnbashi, et al. 2015).

Finally, the ancillary module is a combination of various proteins and domains that are much less common than the core Cas proteins. Cas4 is the exception here being present in several CRISPR loci. Aside from its alleged role in adaptation, this protein is thought to contribute to CRISPR-Cas-coupled programmed cell death (Makarova, Anantharaman, et al. 2012). Other significant components of this module include the Csn2 ATPase, which has been shown to have a role in spacer integration in type II systems, acting as a protecting agent, by accommodating the linear double-stranded DNA in its structure, preventing damage to the DNA (Arslan, et al. 2013).

CRISPR-Cas systems as a defense mechanism

The first experimental evidence that the CRISPR-Cas system is indeed an antiviral defence system was obtained from phage infection experiments of the lactic acid bacterium *Streptococcus thermophiles* (Barrangou, Fremaux, et al. 2007). The authors infected the bacteria and posteriorly screened for adaptation of the CRISPR locus in the surviving bacteria, which revealed that a subpopulation of survivors had acquired new spacer sequences, derived from the genome of the challenging phage. Subsequent deletion of these new spacers resulted in loss of the acquired resistance, signifying the correlation of spacer presence and phage resistance

The exact mechanism by which CRISPR-Cas systems silence extrachromossomal DNA is not yet fully understood, but there is an agreement about the immunological function of CRISPR. Generally, CRISPR-Cas immune systems function in three distinct stages that provided DNA-encoded, RNA-mediated, sequence-specific targeting of exogenous nucleic acids: adaptation, expression and interference.

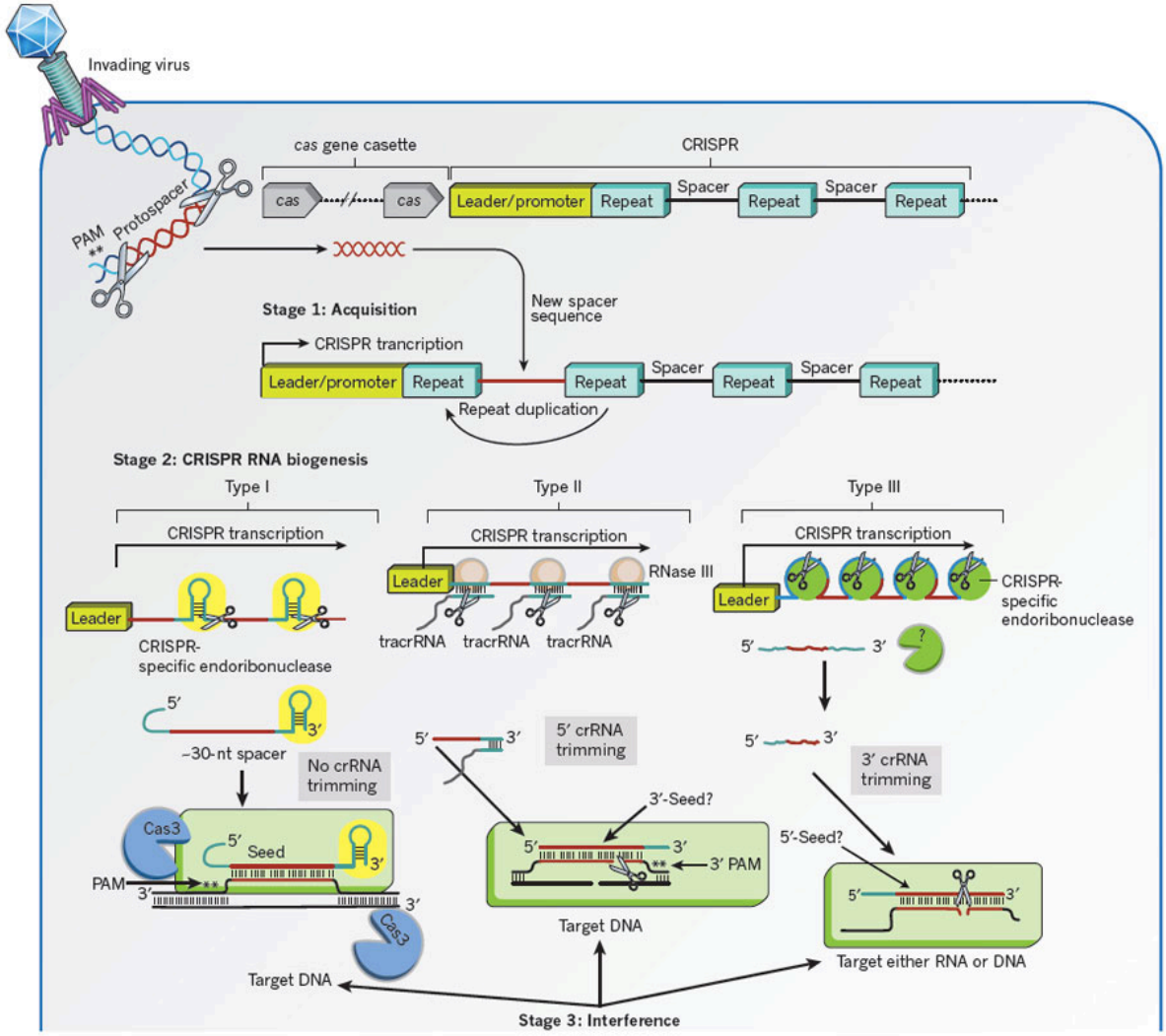


Figure 8 - Schematics of the CRISPR-Cas system immune system; (Stage 1), CRISPR RNA biogenesis (Stage 2) and target interference (Stage 3); (Wiedenheft, Sternberg and Doudna 2012).

Adaptation

The adaptation stage involves the incorporation of fragments of foreign DNA, known as protospacers, from invading viruses and other mobile genetic elements (MGE), into the CRISPR array as new spacers. These spacer sequences provide the sequence memory for a targeted defence against subsequent invasions by the corresponding virus.

In this stage, the role of both Cas1 and Cas2 is fundamental (Nunez, et al. 2014). These Cas proteins, which are present in most known CRISPR-Cas systems, form a complex that represents the adaptation module, being Cas1 the integrase that mediates the acquisition of new spacers. This acquisition, in many CRISPR system types, excepting type III systems, is dependent on the recognition of a short protospacer adjacent motif (PAM) in the target DNA. In type I and type II systems, protospacers are excised at positions adjacent to a PAM sequence, with the other end of the spacer cut using a ruler mechanism inherent to the Cas1 protein, thus maintaining the regularity of the spacer size in the CRISPR cassette (Wiedenheft, Sternberg and Doudna 2012). The protospacer is then ligated to the direct repeat adjacent to the leader sequence, but not exclusively (Erdmann and Garrett 2012). The CRISPR-repeat sequence is duplicated for each new spacer sequence added, thus maintaining the repeat–spacer–repeat architecture.

The short length of PAMs can't provide the level of discrimination to avoid sampling chromosomal DNA. An additional self / non-self mechanism does occur, since the incorporation of spacers from the chromosome is not as common (Richter, Chang and Fineran 2012).

Expression

Spacer acquisition is the first step of immunization, but successful protection from viruses, or other MGEs, challenge, requires the CRISPR to be transcribed and processed into short CRISPR-derived RNAs (crRNAs).

Depending on the CRISPR-Cas system type, this stage of immunization occurs with the aid of different Cas proteins.

For type I and III systems, multiprotein CRISPR RNA (crRNA)-effector complexes mediate the processing. In type I systems, the primary processing of the pre-crRNA is achieved by the Cas6 endoribonuclease within the Cascade complex. Cleavage occurs at the base of the stem-loop formed by the repeat RNA to release mature crRNAs. The Cascade complex recruits the Cas3 nuclease to nick the DNA strand complementary to the protospacer, immediately downstream of the region of interaction with the crRNA spacer (Sikunas, et al. 2013). In type III-A and III-B systems, the protein complexes are respectively known as *csm* and *cmr*. These systems use a Cas6 homolog that does not require hairpin loops in the direct repeat for cleavage. Cas6 cleaves the pre-crRNA to generate intermediate-crRNAs (int-crRNAs) that are incorporated into a *cmr*/Cas10 or *csm*/Cas10 complex, where further maturation occurs through the trimming of 3' end sequences (Hale, et al. 2012).

In type II systems, primary processing requires the annealing of a small *trans*-encoded RNA (tracrRNA) to the repeat sequences of the pre-crRNA. With the guiding of this tracrRNA, cleavage of

the dsRNA, by the conserved endogenous RNAase III, occurs (Deltcheva, et al. 2011). Primary processing happens also with the activity of Cas9, being followed by the trimming of the 5' end repeat and spacer sequences of the int-crRNA to yield mature crRNAs.

Interference

In the final stage of immunization, the guide RNAs, direct Cas endonucleases for sequence-specific targeting, cleavage and degradation of complementary nucleic acids. At this stage, the mature crRNA remains bound to the Cas9 protein or to the multisubunit crRNA effector complex.

For type II systems, target cleavage requires the Cas9 endonuclease to form a complex in combination with the dual guide RNA, crRNA and tracrRNA, and the target dsDNA. After binding to the PAM motif, RuvC-like nuclease and HNH nuclease domains of Cas9, each of which cleaves one of the strands of the target protospacer region, 3 nucleotides upstream of the PAM (Jinek , et al. 2012).

During the interference stage, in type I systems, the PAM sequence is recognized on the crRNA-complementary strand and is required along with crRNA annealing. A conformational change in the Cascade complex, caused by the correct base pairing between the crRNA and the protospacer, results in the recruitment of Cas3 for DNA degradation.

In type III systems, genetic and biochemical evidence indicates that type III-A and III-B systems can cleave DNA (Hatoum-Aslan, Maniv, et al. 2014) and RNA (Hale, et al. 2012), respectively, and work independently of PAMs. The crRNAs associated with the Cas RAMP module (*cmr*) effector complex, which cleaves targeted RNAs, contain an 8-nucleotide 5' sequence tag that is critical for crRNA function.

The mechanism for distinguishing self from foreign DNA during interference is built into the crRNAs and is therefore inferred to be common to all three systems. For DNA-targeting type I and type II systems, the requirement for a PAM sequence in the invader avoids damage to the host genome at the site of the corresponding guide sequence in the CRISPR cassette, since PAMs are not found in the flanking repeat sequences. In contrast, for the DNA-targeting type III-A and RNA-targeting *cmr* complex, where host genome damage is not a risk, target DNAs, or RNAs, containing sequences with perfect homology to the 5' tag of the crRNA are efficiently cleaved (Hale, et al. 2012) (Hatoum-Aslan, Palmer, et al. 2013).

Phage adaptation to CRISPR

Being the most populous biological entities on Earth, phages have evolved, along side CRISPR systems, developing mechanisms to escape or inhibit CRISPR activity. So far, there are two known mechanisms for phages to evade the CRISPR-Cas system: mutation and activity of "anti-CRISPR genes". The first is more of a low-frequency event, consisting on single or multiple mutations in the protospacer's sequence or in the conserved protospacer-adjacent motif. Mutations in the PAM sequence eliminate CRISPR-mediated immunity even in the presence of a perfect spacer-protospacer match. On the other hand, it has been shown that certain protospacer related mutations do not lead to phage escape, implying a relaxed specificity of the CRISPR-Cas system regarding the requirements for crRNA matching, especially in regions further from the protospacer motif. In fact, in the case of *Escherichia coli* subtype CRISPR-Cas system, the requirements for crRNA matching are strict only for a 7 bp seed region of a protospacer immediately following the essential PAM (Semenova, et al. 2011).

More recent is the discovery of "CRISPR inhibitors" that reside in certain phages (Bondy-Denomy, et al. 2013). Phage genes have been discovered that can neutralize most of the prevalent bacterial anti-phage defences, and a functional CRISPR-Cas system is no exception. Bondy-Denomy and colleagues proved the existence of eight anti-CRISPR genes, in the genomes of bacteriophages infecting *Pseudomonas aeruginosa,* predicted to encode five different proteins. Tests with three different bacteriophages have produced some interesting results, namely that the anti-CRISPR genes apparently do not function through protection of specific DNA sequences on the phages, exerting their effects at a step occurring after formation of the crRNA-Cas complex (not affecting the biogenesis of either the crRNA or the Cas protein). This phage machinery is a necessity and, at the same time, a sufficiency for the annulment of the CRISPR-Cas system bacterial immunity (Bondy-Denomy, et al. 2013). Interestingly enough, the occurrence of self-targeting spacers in some CRISPR arrays may be associated with these anti-CRISPR genes (Louwen, et al. 2014), which will act as autoimmunity.

It is postulated that the existence of anti-CRISPR genes may be one of the explanations for phages continued proliferation despite the undeniable potency and efficiency of the CRISPR-Cas machinery. In the same way, their diverseness among phages and other mobile genetic elements may affect the large diversity of CRISPR-Cas systems within bacterial strains, which co-evolve alongside the diversification of anti-CRISPR genes.

Current and Future applications

The discovery of CRISPR-type repetitive elements in the 80's resulted in an intriguing research area of bacterial genetics, with several discoveries covering defence against viruses, and other genetic elements, as well as regulation of gene expression. The range of potential applications for CRISPR-Cas systems goes from protecting bacteria against viral infections, as studied, to genome editing of microbial and mammalian cells, which presents itself as a possible future of gene therapy approaches and gene editing.

Genome editing

Notwithstanding the natural immune role of CRISPR-Cas systems in prokaryotes, the most recent research as been focusing on using the machinery for genome editing and transcriptional control in eukaryotes.

Cas9 endonuclease became a hot topic among researchers, after being discovered its mechanism of DNA cleavage. The Cas9 endonuclease can be programmed with guide RNA engineered as a single transcript to target and cleave any dsDNA sequence of interest.

RNA-programmed Cas9 offers a powerful alternative to existent genome manipulating techniques using artificial enzymes like Zinc-Finger Nucleases (ZFNs) and Transcription-Activator Like Effector Nucleases (TALENs). The latter both rely on protein-mediated recognition of the DNA, which means that every new target requires the engineering of new proteins. In contrast to these techniques, Cas9 relies on complementary base pairing and protein mediated recognition of an adjacent short sequence motif (PAM). Alongside the diversity of Cas9 proteins and the simplicity of RNA-guided programming, the need for sophisticated protein engineering is revoked, and affords rapid generation of designer nucleases (van Erp, et al. 2015).

Strain typing

Bacterial strain typing, or identifying bacteria at the strain level, is particularly important for diagnosis, treatment, and epidemiological surveillance of bacterial infections. Beyond strain identification, high-resolution subtyping methods can provide opportunities to improve the understanding of bacterial population genetics and dynamics, epidemiology and evolution.

The ultimate genotyping is the determination of the complete genome sequence of an organism. Otherwise, specific genetic polymorphisms can be used to compare strains, such as presence/absence of insertion elements (IS), single nucleotide polymorphisms (SNP), and variable number of tandem repeats (VNTR) (Pourcel and Drevet 2013). Nowadays, several high-throughput and high-resolution typing methods are used, including PCR-centered approaches such as VNTR analysis and multilocus sequence typing (MLST), and plus, more recently, whole-genome sequence-based techniques. For an extensive review on typing methods, see (Sabat, et al. 2013).

CRISPRs' loci dynamic nature allowed them to be rendered as ideal targets for molecular subtyping. Although it can't be considered the sole source of genetic variability for bacterial genotyping, CRISPR

polymorphism is elevated in some species, and can be exploited for rapid genotyping of even closely related strains.

Differences between strains containing CRISPRs may occur at the spacer level, with deletions and insertions, or with SNPs introduced into the spacers of the direct repeats. All these changes contribute to strain-to-strain differences. For example, in a CRISPR array, the opposite end to the leader can contain spacers conserved across various strains and can be useful to cluster related phylogenetic group of strains. Based in the spacer variability existent, a technique called spacer-oligotyping (spoligotyping) was developed. In this method, probes for specific spacers are bound to a membrane and hybridization patterns of labelled PCR products, primed from the CRISPR repeats, are measured. This technique is a useful approach for phylogenetic studies but it has a limited value for evolutionary studies, since it only determines the presence of pre-established spacers (Comas, et al. 2009). Nevertheless, this has been the number one technique for a number of species including the *Mycobacterium tuberculosis* complex (MTBC) (Streicher, et al. 2007). Automation and the use of microbeads to replace the membrane represent some improvements to classical spoligotyping, enabling a higher-throughput procedure (Shariat and Dudley 2014)

Next generation sequencing brought some advances in CRISPR related typing techniques allowing for the rise of sequence-based methods. With these methods, the entire CRISPR spacer arrays are PCR amplified and sequenced. The majority of sequence-based typing has been done in *Salmonella* (Shariat and Dudley 2014).

So far, even considering some limitations, CRISPR-Cas systems were found to be a useful tool for typing bacterial diversity, especially when in combination with other typing techniques like amplified length polymorphisms (AFLP) and multilocus sequence typing (MLST) (Louwen, et al. 2014).

Engineered defence against viruses

Phage infection is a very serious problem for some industries that rely on bacteria, such as the dairy and wine industries, which spend heavily in efforts to combat this natural occurrence (Sturino and Klaenhammer 2006). Making use of the efficiency of CRISPR-Cas systems for battling phage invasion might offer a partial solution to reduce the high costs associated with phage-mediated culture losses. By engineering the CRISPR array, adding spacers derived from conserved regions of known infecting phages, could result in the boost of the culture's immunity.

The company Danisco (DuPont) was an initial pioneer of commercial use of CRISPR technology to enhance viral immunity in bacteria used to make yogurts and cheese (van Erp, et al. 2015).

Tools for CRISPR detection and analysis

Detection software

Several software applications are available for identifying various forms of repeats, like Patscan (Dsouza, Larsen and Overbeek 1997), REPuter (Kurtz, et al. 1999) or Pygram (Durand, et al. 2006), however, used with the objective of finding CRISPRs, they usually require considerable manual post-processing. The growing interest in CRISPRs has led to the development of different bioinformatic tools, software and web resources, for the sole purpose of the automatic detection and analysis of CRISPR systems. A few specific programs were conceived for this purpose of identifying CRISPR arrays in genomes or other DNA sequences, the most used being PILER-CR (Edgar 2007), CRISPR Recognition Tool (CRT; (Bland , et al. 2007)) and CRISPRFinder (Grissa, Vergnaud and Pourcel 2007). The first two are software tools: PILER-CR needs to be compiled before use and CRT requires either to install JRE (Java Runtime Environment) or compile the source files. In comparison, CRISPR Finder presents itself as a web server tool.

All three detection tools mentioned are of public domain, only accept the files in FASTA format, and apply post-processing filters to separate real CRISPR arrays from false predictions.

PILER-CR

As mentioned, PILER-CR is a dedicated software tool for the identification and preliminary analysis of CRISPR arrays, with a greater focus on the array repeats. The program algorithm is designed to identify the characteristic signature of CRISPR repeats, using alignment matrices, being its goal to find a chain of local alignments that meets the criteria for being a CRISPR array, i. e., repeats and spacers are within the expected ranges of length and sequence conservation. These ranges constitute the most important parameters of the PILER-CR algorithm: the *minrepeat*[4] parameter, minimum length of a repeat, is set as default to 16 and *maxrepeat*[4]*,* maximum length of a repeat, is set for 64. Spacer minimum length, parameter *minspacer*[4]*,* is set for 8, and spacer maximum length, *maxspacer*[4]*,* is also, per default, 64.

Also worth mentioning, are other major parameters of the algorithm, which can also be changed by the user: *minarray*[4] represents the value for the minimum number of repeats in an array, being 3 the default number; the algorithm author describes the parameters *minrepeatratio*[4] (default value is 0.9) and *minspacerratio*[4] (default value is 0.75) as the ratio of the smallest to the longest repeat/spacer sequence, thus a value close to 1.0 requires lengths to be similar, whereas 1.0 means identical lengths. Spacer's lengths on occasion vary significantly, so the default ratio is smaller than for repeat lengths.

[4] Command-line option denomination for PILER-CR program.

The program output is designed to facilitate manual analysis of the reported CRISPR arrays, showing a multiple alignment of the repeats in each array, obtained using fast options of algorithm MUSCLE[5], as well as flanking regions and spacer's sequences. The algorithm runs with a time and space complexities of O (L^3), where L is the input sequence length. Processing speed, sensitivity and specificity were shown to be high (Edgar 2007) (Bland , et al. 2007).

The software can accessed online at http://drive5.com/pilercr.

CRISPR Recognition Tool (CRT)

Also a CRISPR detection tool, CRT uses a simple sequential search technique that detects repeats directly from a DNA sequence. This technique requires no major conversion or reprocessing of the input. The algorithm basis is that it tries to find a series of short exact repeats of length *k* that are separated by a similar distance and then extending those *k*-mer matches to the actual repeat length. Once actual repeats are found, they are filtered to remove those that do not meet CRISPR specific requirements, meaning, the repeat length should be between the defined limits and spacers are non-repeating and similarly sized. These length limits are defined as parameters by the algorithm, *minRL*[6] and *maxRL*[6], representing the minimum and maximum length of a CRISPR repeated region (default values are 19 and 38, respectively). Also important parameters are: *minNR*[6], minimum number of repeats a CRISPR must contain (default is 3, similar to PILER-CR algorithm); *minSL*[6] and *maxSL*[6], minimum and maximum length of a CRISPR's non-repeated region, or spacer region (default values are 19 and 48, respectively). The user can change all the parameters.

The program output, similarly to PILER-CR, facilitates the user post analysis of CRISPR arrays, showing flanking regions and repeats and spacers sequences. The algorithm runs with a time complexity of O (*nm/l*), where *n* is the length of the input sequence, *m* is the length of the search range and *l* is the interval at which the search window advances. The algorithm is also linear in space, since repeats are detected directly from the input sequence. Processing speed, sensitivity and specificity were also shown to be high (Bland , et al. 2007).

The software can accessed online at http://www.room220.com/crt/.

[5] MUSCLE (multiple sequence comparison by log-expectation) is a public domain, multiple sequence alignment software, for protein and nucleotide sequences. MUSCLE is freely available at http://www.drive5.com/muscle. (Edgar, R. C.. 2004)

[6] Command-line option denomination for CRT program.

Spacer analysis tool

CRISPR Target

This tool, designed to detect, and interactively explore, the targets of CRISPR RNA spacers, was published and made freely available in 2013 (Biswas, et al. 2013). Being the first of this kind specially designed for the purpose, it allows discovering targets in newly sequenced genomic or metagenomic data.

Users can provide their input as either spacers in FASTA format, or as CRISPRFinder, PILER-CR, or CRISPR Recognition Tool (CRT) output files. Putative protospacer targets can be identified, following a BLAST N search of the spacer input against a number of databases or user-uploaded sequences. These databases include ACLAME genes, GenBank-nt, GenBank-Environmental, GenBank-Phage, RefSeq-Microbial, RefSeq-Plasmid, RefSeq-Viral and parts of CAMERA. The default settings are made to allow the sensitive detection of potential targets, but users have the ability to modify the search parameters, such as e-value, word size and penalties for gaps and match/mismatch. The default values: *gap open* and *extend* are -10 and -2; *Nucleotide score* is +1 for a match and -1 for a mismatch; *word size* is 7 and *e-value* is 1. BLAST calculates the scores over the length of the match, and only shows this match. For example, a spacer of 32 bases that matches to a target in 17 of 20 bases would score 20-3=17 and 20 bases would be the output.

The output provided by the algorithm is either visual in HTML format, or can also be saved as text. The target sequence, protospacer, is typically displayed as an R-loop (showing both orientations, 3' to 5' and 5' to 3'), depicting a specified part of the crRNA, as well as both the target and non-target strand of the double-stranded target DNA (Biswas, et al. 2013). All putative spacer/protospacer targets passing the BLAST screen are displayed.

The tool can be accessed online at http://bioanalysis.otago.ac.nz/CRISPRTarget.

A CRISPR publicly available database

A public database for CRISPR, named in short CRISPRdb, was created in 2007, and can be queried online at http://crispr.u-psud.fr/crispr. Up to date, the database contains 4065 CRISPRs found, from 2762 analysed genomes of Bacteria and Archaea.

The core application consists of two main programs: CRISPRFinder (to construct the database) which identifies CRISPRs and extracts the repeated and unique sequences from a genomic sequence, and Database Tools for downloading prokaryotic genomes from the NCBI ftp site, saving CRISPRs and making updates, making it possible to keep the database up to date automatically (Grissa, Vergnaud and Pourcel 2007). The CRISPRFinder program can also be ran interactively through the website for any user submitted sequences under 67,000,000 bp The identification of putative CRISPRs starts with the finding of the largest number of possible CRISPRs, specially the short ones, containing one or two spacers. The main idea is to first find possible CRISPR localizations and then verify if these regions contain a cluster that has the features of recognizable CRISPR, meaning, it contains at least three repeats. This step is done using Vmatch package (Vmatch n.d.) in order to detect maximal repeats (i. e. a repeat that can't be extended in any direction without incurring a mismatch). At this point, reported matches must obey two important parameters: the repeats must be between 23 to 55bp, with one possible mismatch, and the gap between them must be within 25 to 60bp. The clusters that insert themselves in these parameters are then submitted to a filtering process designed to find what the authors denominate as "confirmed CRISPRs". The latter contain DR conserved at least to 80%, spacers are within the range of 0.6x and 2.5x the DR length and are no more than 60% similar between them or the DR. Algorithm processing speed, sensitivity and specificity were also proved to be high, comparing to other CRISPR detection tools (Grissa, Vergnaud and Pourcel 2007).

CRISPRdb provides another set of available tools that allow the user further analysis of CRISPR arrays. The BLAST CRISPR tool (BLAST CRISPR n.d.) is of use when trying to validate a questionable CRISPR. From this page, a candidate DR region (or spacer) can be compared to all DRs (or spacers) characterised so far from CRISPR structures present in the database. The FlankAlign tool (FlankAlign n.d.) is useful to compare CRISPR flanking sequences. And finally, CRISPRtionary (CRISPRtionary n.d.) allows to create a dictionary containing the spacers identified in the submitted sequences and comparing the spacer's arrangement.

Tools to further analyse and classify CRISPR arrays

CRISPRmap

CRISPRmap presents itself as the most recent CRISPR related tool (Lange, et al. 2013). It consists of an automated classification system of CRISPR repeats, providing a comprehensive analysis of CRISPR structure and sequence conservation based on the largest data set of repeat sequences available up to date. Characteristics include cleavage sites, patterns of RNA structure motifs and sequence conservation. CRISPRmap approach considers not only the pairwise similarities between repeats, but also how the binding affinity of Cas proteins is affected by the repeat structure, namely, a small hairpin structure (a key binding motif for Cas endoribonucleases in several systems). The classification system provides a more complete insight into the diversity of CRISPR systems.

Working as a web server tool, CRISPRmap accepts as an input properly identified repeats, in FASTA format. The output result consists of the clustering of the entered repeats to one of the six superclasses, and attribution, if possible, to a structure motif and/or a sequence family. First, if one of the repeat sequences already exists in the CRISPRmap cluster tree it is automatically assigned to the corresponding structure motif and/or sequence family. If that's not the case, the next step consists of the use of a RNA structure prediction algorithm, RNAfold (Hofacker and Stadler 2006), to determine whether the repeat sequence is structured or unstructured. If the minimum free energy structure is the unstructured sequence, i.e. contains no bp, it remains unassigned to a structure motif. And although a structure is predicted, the repeat does not necessarily belong to a conserved structure motif. The repeat sequence is added to all groups of repeats already assigned to one of the structure motifs and RNAclust tool (Will, Reiche, et al. 2007) is re-runned. If the repeat falls into or next to one of the existing structure motifs, it is assigned to the motif by firstly being folded by RNAfold (Hofacker and Stadler 2006) with the calculation of a structure dotplot. This dotplot is aligned with the consensus dotplot of the structure motif using LocARNA (Hofacker and Stadler 2006). The repeat is then assigned to the motif if it is able to fold into the consensus structure of that respective motif with at most one bp missing (Lange, et al. 2013). Attribution of the repeat to a conserved sequence family is done by comparing it to the previously calculated ClustalW (Thompson, Higgins and Gibson 1994) sequence profiles. If the "new" repeat is sufficiently similar it's added to the family (Lange, et al. 2013). Lastly, with a final run of RNAclust (Will, Reiche, et al. 2007) on all repeat sequences, CRISPRmap cluster tree is updated and input sequences locations are highlighted. So far, the authors have analysed 4719 consensus repeat sequences covering 24 families and 18 structural motifs. The six superclasses were constructed based on sequence-and-structure similarities and tree topology. They are labelled from A to F, and are ordered according to generally decreasing conservation. Superclass A contains highly conserved CRISPRs repeats on the sequence level, but only a few small structure motifs. Superclasses B–C contain sequence families that roughly correspond to one structure motif each; the same is true for half of superclass D. The other half of superclass D and superclass E contain little sequence conservation, but many small conserved motifs. The bacterial repeats in superclass F are divergent (Lange, et al. 2013).

CRISPRs in Metagenomic studies

As a reflection of the infectious dynamics of microbial communities, the study of CRISPRs is an essential complement to the study of the human microbiome, comprehending both disease ecology and ecological immunity. Infectious disease works to maintain both species diversity and genotypic diversity within a species. As such, infectious agents may be at least partially responsible for the amazing species diversity and turnover found throughout the human microbiome. The ability of CRISPRs to prevent plasmid spread is medically relevant, in that the exchange of conjugative elements is perhaps the dominant mechanism by which antibiotic resistance genes move within a biome, and by which pathogens acquire resistance. CRISPR activities could be expected to retard this exchange.

The evolution of sequencing techniques permitted the “rise” of metagenomics, which allowed one to obtain a complete snapshot of coexisting microorganisms, both prokaryotes and their viruses, in several ecosystems. To date, CRISPR systems have been analysed in metagenomic datasets of several environments including the ocean metagenome produced by the Global Ocean Sampling (GOS) expedition (Sorokin, Gelfand and Artamonova 2010), acidic hot springs in Yellowstone National Park (Bolduc, et al. 2012), the bovine rumen microbiome (Berg Miller, et al. 2012) and Australian hypersaline Lake Tyrrell (Emerson, et al. 2013).

The increasing availability of shotgun metagenomic datasets from the HMP, MetaHIT and other projects, enabled the further exploration of the distribution and diversity of CRISPR arrays in the human microbiome, allowing the possible discovery of new arrays, across different body sites. Indeed, CRISPR arrays and CRISPR-Cas systems have been characterized across different individuals and body sites through independent projects (Pride, et al. 2011) (Robles-Sikisaka, et al. 2013), with a particular attention to the gut metagenome (Stern, Mick, et al. 2012) (Gogleva, Gelfand and Artamonova 2014), and as part of HMP (Rho, et al. 2012).

Different approaches have been used for studying CRISPR(s): some studies focused more on the *de novo* prediction of CRISPR cassettes using the whole metagenome assemblies, and other, focused on a targeted assembly, making more use of the raw reads. In the latter, raw reads containing CRISPR repeats were collected, followed either by a more comprehensive analysis of spacer content (Stern, Mick, et al. 2012) or specific reassembly of repeat-containing reads into CRISPR contigs (Rho, et al. 2012) with the purpose of studying their evolutionary dynamics.

Project Aim

This project was designed with the objective of further exploring the complex human gut ecosystem and the bacteria in it existent, with a special focus to their defence systems and, in particular, the adaptive immunization provided by the CRISPR-Cas systems.

Materials and Methods

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 (J. Qin, Y. Li, et al. 2012). The sequenced and analyzed data from the study can be freely accessed online at GigaScience database (GigaDB) (Li, et al. 2012).

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 (J. Qin, Y. Li, et al. 2012). 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**.

Identification and analysis of CRISPR cassettes

Detection of CRISPR cassettes

To construct a set of CRISPR cassettes for the metagenomic dataset two freely available CRISPR-finding algorithms were used, PILER-CR and CRISPR Recognition Tool (CRT), 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

Filtering step

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.

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+).

Identification of unique repeats and construction of a collection of non-redundant repeat sequences

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, which 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.

Significant Clusters

Significant clusters are defined as redundant repeat sequences that contain six or more recurrences in different CRISPR cassettes.

The collection of CRISPRs associated with these sequences was made manually.

Comparison to the CRISPR database

Each unique repeat sequence was matched against the CRISPR database, CRISPRdb (accessible at http://crispr.u-psud.fr/crispr/BLAST/CRISPRsBlast.php), 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.

CRISPRmap

All repeats from the non-redundant set were analysed with the CRISPRmap tool (Lange, et al. 2013). 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.

Taxonomy of metagenomic contigs containing CRISPR cassettes

To determine contig taxonomy, contigs were subjected to a BLASTX (version 2.2.32) (Altschul, et al. 1997) search against the non-redundant protein collection (Benson, et al. 2013), 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.

CRISPR-associated proteins

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 (Grissa, Vergnaud and Pourcel 2007) 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.

Identification of protospacers

In order to identify and label spacer origin, CRISPRTarget program was used (Biswas, et al. 2013). 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.

crAssphage

To infer about the presence of a largely common bacteriophage, Gut phage BED-2012 or crAssphage (Dutilh, et al. 2014), 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.

Results and discussion

Metagenomic dataset

The human gut microbiota is one of the most complicated microbial ecosystems in the human body and has important associations with human health. Alongside with its title for having the greatest diversity of microorganisms in the human body, the human gut is considered to be one of the most interesting objects of research.

As a considerable effort directed to large-scale investigation of the human microbiome, big projects like HMP and MetaHIT have conducted studies intended to understand the role of the endogenous flora in health and disease, and made publicly available their metagenomic data. In this scope, several projects have been undergoing, and completed, studying and characterizing CRISPR systems in these metagenomic datasets of several human populations with different geographical backgrounds. So far, CRISPR has been analysed in populations from North and Mediterranean European countries, North America and Japan.

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 (J. Qin, Y. Li, et al. 2012). 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.

Detection and Characterization of CRISPR cassettes

CRISPR Detection software

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 (Bland , et al. 2007). 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 (Rho, et al. 2012). 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.

```
CRISPR 32   Range: 59934609 - 59935555
POSITION    REPEAT                SPACER
--------    --------------------------------    --------------------------------
59934609        AAGGCTTGATGTTGTATCATAAGGAGGATGAC    AGCTCCATCTAAATTTATCGAACAAGCTATACC    [ 32, 33 ]
59934674        ATTTCAACTCACATCCTCACAAGGAGGATGAC    CTCGTTCTGCTGCTTTCTCATCCCAGCTGTATGACT     [ 32, 36 ]
59934742        ATTTCAACTCACATCCTCACAAGGAGGATGAC    TCTATTTCTGTGCCGGCACTGGACAAGATATAT    [ 32, 33 ]
59934807        ATTTCAACTCACATCCTCACAAGGAGGATGAC    AGATGTTTTATAAGTGTGGTTGATAAGCATTTCCT [ 32, 35 ]
59934874        ATTTCAACTCACATCCTCACAAGGAGGATGAC    TAC>SCAFF0LD40528_4_2  LENGT     [ 32, 28 ]
59934934        H=670AACTCACATCCTCACAAGGAGGATGAC    TACTGATTCTCATAGTTGTTCTTATCGCAAGAT    [ 32, 33 ]
59934999        ATTTCAACTCACATCCTCACAAGGAGGATGAC    GATAAATACTATTAGTATGAAAATATTATTAGAT   [ 32, 34 ]
59935065        ATTTCAACTCACATCCTCACAAGGAGGATGAC    GTTGGGTGGTCTTAGAGTAACTTATGATGATGGC   [ 32, 34 ]
59935131        ATTTCAACTCACATCCTCACAAGGAGGATGAC    CGCCGGAAACGGTCTACTACATACGAGAAGTGAT   [ 32, 34 ]
59935197        ATTTCAACTCACATCCTCACAAGGAGGATGAC    TTTGATTTTTGCAGGAGCATTTGACCGATTTAC    [ 32, 33 ]
59935262        ATTTCAACTCACATCCTCACAAGGAGGATGAC    AGGAAGTACATCAGAACTTGGTAATGGAAGTGGT   [ 32, 34 ]
59935328        ATTTCAACTCACATCCTCACAAGGAGGATGAC    GGAGCATACAATGCAAGATATATTATCAGTAGT    [ 32, 33 ]
59935393        ATTTCAACTCACATCCTCACAAGGAGGATGAC    AGGCTTTTTCTTTTCTAAGCATAGACATAGTTAT   [ 32, 34 ]
59935459        ATTTCAACTCACATCCTCACAAGGAGGATGAC    AACACTTCAACTTTAACATGTACACTATCTTAT    [ 32, 33 ]
59935524        ATTTCAACTCACATCCTCACAAGGAGGATGAC
--------    --------------------------------    --------------------------------
Repeats: 15 Average Length: 32      Average Length: 33
```

Figure 9 – Example of an output from CRT showing a predicted CRISPR cassette between two different contigs, evidencing the problem with CRT regarding the non individualization of the contigs.

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 (Skennerton, Imelfort and Tyson 2013). 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 of the CRISPRs in the metagenomic dataset. 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 (Skennerton, Imelfort and Tyson 2013).

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.

Filtering step

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 (Gogleva, Gelfand and Artamonova 2014) 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.

Predicted CRISPR cassettes

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.

The full set of identified cassettes is shown and characterized in **Table 2** and **Supplemental File 2**.

Table 2 – Characteristics of identified CRISPR cassettes and spacers by both PILER-CR and CRT, for the metagenomic dataset relative to the diabetic individuals and the healthy individuals.

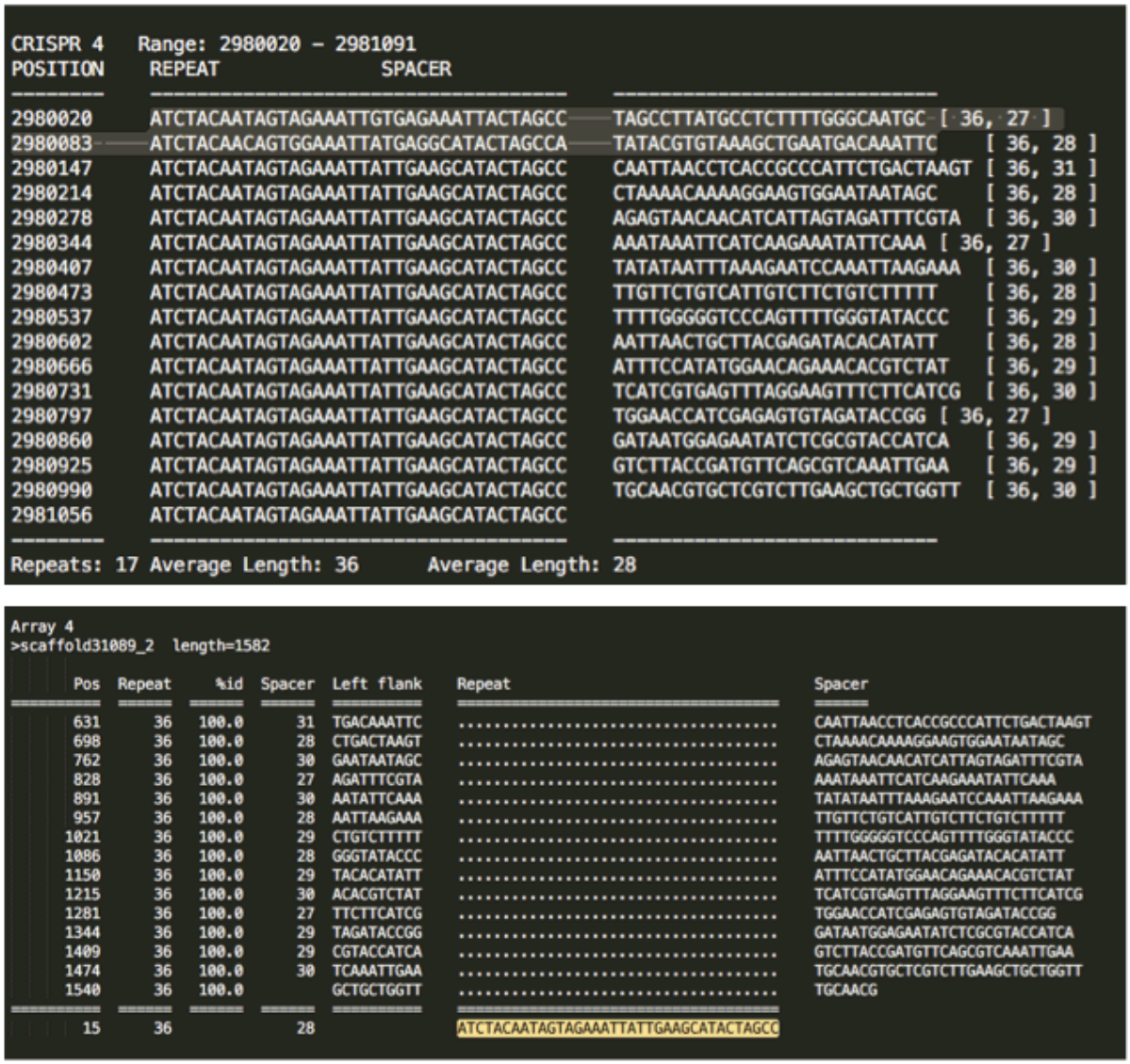
	Metagenomic dataset group	Type-2 diabetic individuals	Healthy (non-diabetic) individuals
	Cassettes		
Identified by PILER-CR and CRT		630	695
Average Length (in bp)		882	786
Average CRISPR per individual		9	10
	Spacers		
Total number		8,140	17,739
	Repeats		
Average size (in bp)		32	32

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 10).

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+).



```
CRISPR 4   Range: 2980020 - 2981091
POSITION    REPEAT                  SPACER
--------    ------------------------------------    ----------------------------
2980020     ATCTACAATAGTAGAAATTGTGAGAAATTACTAGCC    TAGCCTTATGCCTCTTTTGGGCAATGC [ 36, 27 ]
2980083     ATCTACAACAGTGGAAATTATGAGGCATACTAGCCA    TATACGTGTAAAGCTGAATGACAAATTC    [ 36, 28 ]
2980147     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    CAATTAACCTCACCGCCCATTCTGACTAAGT [ 36, 31 ]
2980214     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    CTAAAACAAAAGGAAGTGGAATAATAGC    [ 36, 28 ]
2980278     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    AGAGTAACAACATCATTAGTAGATTTCGTA  [ 36, 30 ]
2980344     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    AAATAAATTCATCAAGAAATATTCAAA [ 36, 27 ]
2980407     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    TATATAATTTAAAGAATCCAAATTAAGAAA  [ 36, 30 ]
2980473     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    TTGTTCTGTCATTGTCTTCTGTCTTTTT    [ 36, 28 ]
2980537     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    TTTTGGGGGTCCCAGTTTTGGGTATACCC   [ 36, 29 ]
2980602     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    AATTAACTGCTTACGAGATACACATATT    [ 36, 28 ]
2980666     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    ATTTCCATATGGAACAGAAACACGTCTAT   [ 36, 29 ]
2980731     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    TCATCGTGAGTTTAGGAAGTTTCTTCATCG  [ 36, 30 ]
2980797     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    TGGAACCATCGAGAGTGTAGATACCGG [ 36, 27 ]
2980860     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    GATAATGGAGAATATCTCGCGTACCATCA   [ 36, 29 ]
2980925     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    GTCTTACCGATGTTCAGCGTCAAATTGAA   [ 36, 29 ]
2980990     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC    TGCAACGTGCTCGTCTTGAAGCTGCTGGTT  [ 36, 30 ]
2981056     ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC
--------    ------------------------------------    ----------------------------
Repeats: 17 Average Length: 36      Average Length: 28
```

```
Array 4
>scaffold31089_2  length=1582

      Pos  Repeat     %id  Spacer  Left flank    Repeat                                Spacer
==========  ======  ======  ======  ==========  ====================================  ======
      631      36   100.0      31  TGACAAATTC  ....................................  CAATTAACCTCACCGCCCATTCTGACTAAGT
      698      36   100.0      28  CTGACTAAGT  ....................................  CTAAAACAAAAGGAAGTGGAATAATAGC
      762      36   100.0      30  GAATAATAGC  ....................................  AGAGTAACAACATCATTAGTAGATTTCGTA
      828      36   100.0      27  AGATTTCGTA  ....................................  AAATAAATTCATCAAGAAATATTCAAA
      891      36   100.0      30  AATATTCAAA  ....................................  TATATAATTTAAAGAATCCAAATTAAGAAA
      957      36   100.0      28  AATTAAGAAA  ....................................  TTGTTCTGTCATTGTCTTCTGTCTTTTT
     1021      36   100.0      29  CTGTCTTTTT  ....................................  TTTTGGGGGTCCCAGTTTTGGGTATACCC
     1086      36   100.0      28  GGGTATACCC  ....................................  AATTAACTGCTTACGAGATACACATATT
     1150      36   100.0      29  TACACATATT  ....................................  ATTTCCATATGGAACAGAAACACGTCTAT
     1215      36   100.0      30  ACACGTCTAT  ....................................  TCATCGTGAGTTTAGGAAGTTTCTTCATCG
     1281      36   100.0      27  TTCTTCATCG  ....................................  TGGAACCATCGAGAGTGTAGATACCGG
     1344      36   100.0      29  TAGATACCGG  ....................................  GATAATGGAGAATATCTCGCGTACCATCA
     1409      36   100.0      29  CGTACCATCA  ....................................  GTCTTACCGATGTTCAGCGTCAAATTGAA
     1474      36   100.0      30  TCAAATTGAA  ....................................  TGCAACGTGCTCGTCTTGAAGCTGCTGGTT
     1540      36   100.0          GCTGCTGGTT  ....................................  TGCAACG
==========  ======  ======  ======  ==========  ====================================
       15      36              28              ATCTACAATAGTAGAAATTATTGAAGCATACTAGCC
```

Figure 10 - The case where it is possible to observe that both software's output provides a 100% identical repeat sequence but the number of spacers predicted for the cassette differentiates in more than two units; Output from CRT (on the top) detects an array of 10 repeats and 9 spacers, while the output from PILER-CR (on the bottom) detects 8 repeats and 7 spacer sequences.

Unique repeats

To the set of reliable CRISPR cassettes from the T2D+ dataset correspond 630 repeats (the same number of CRISPR cassettes). From these, we identified 234 unique repeats (about 40% of the total number of repeats). By unique repeats we mean repeats that are not repeated in the dataset and hence, unique for a determined CRISPR cassette and individual. Nonetheless, from the 396 non-unique repeats, we identify 128 unique sequences. In the end, we have a set of non-redundant 362 repeat sequences, from which 65% are indeed unique. See **Supplemental File 3**.

The repeat length fell within the expected range according to the literature, being the average length for a repeat 32 bp (Makarova, Haft, et al. 2011). Also, although there was no computational validating, there is an overall visual agreement about repeat sequence diversity.

The similarity in the repeats of two different CRISPR cassettes might indicate a possible horizontal gene transfer between the strains (Makarova, Haft, et al. 2011). This horizontal gene transfer may be mediated by plasmids, mega plasmids and even prophages that carry the CRISPR (Godde and Bickerton 2006).

The output retrieved from CRISPR database

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 92 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 (Sun and Chang 2014). The last two species are bacteria that are part of the colonic microbiomes responsible for the fermentation of hexose and pentose sugars (Russell, et al. 2013).

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**.

Significant repeat clusters

As explained in "Materials and Methods", the set of non-unique repeats was analysed and grouped according to the shared identical repeat sequence, which resulted in 128 '"clusters" named A to DY. On average, each repeat sequence, is present in 3 different CRISPR containing contigs. But, there are some repeats that are present in more than 5 different CRISPR cassettes, and those are considered "significant clusters".

10 clusters were collected with the ID: E, F, K, L, BD, BN, CB, CE, CQ and DC. See **Supplemental File 5**.

The presence of repeated sequences may indicate one of two things: if the same individual has two identical CRISPR repeats shared by two or more cassettes, that also share a spacer, it could point to a horizontal transfer phenomenon, meaning that part of a cassette might have been acquired by one of the CRISPR arrays; another hypothesis is that the CRISPR cassette belongs to the same bacterial strain.

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".

Contigs containing unique repeats

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.

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.

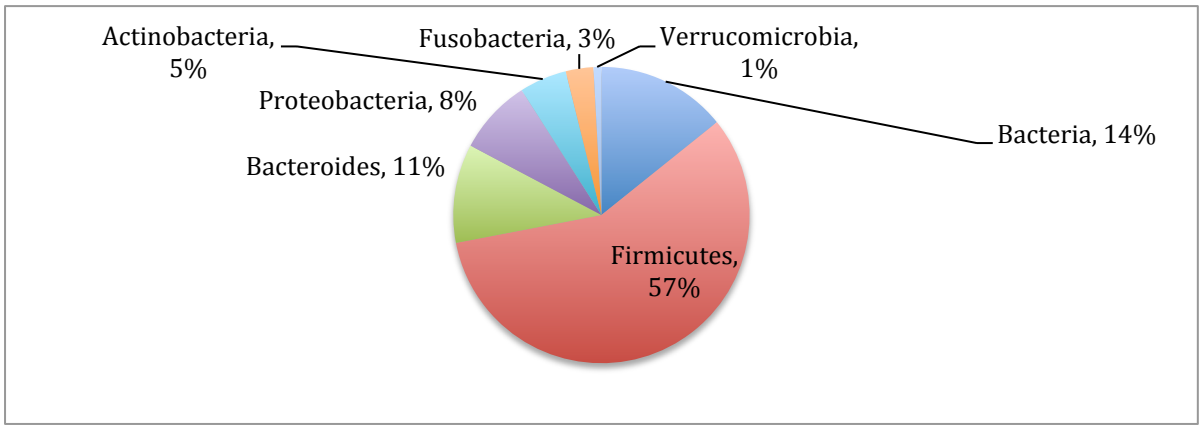


Figure 11 - Taxonomy of CRISPR-containing contigs with unique repeats; Firmicutes has the largest fraction attributed, counting for more than half of the analysed contigs; Bacteria is referent to the fraction of contigs to which wasn't possible to attribute a taxonomy, so it was attributed this generic label.

The 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 (Cho and Blaser 2012).

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 (Tremaroli and Bäckhed 2012).

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 (J. Qin, Y. Li, et al. 2012), 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.

CRISPRdb taxonomy VS Significant clusters taxonomy

The CRISPR-containing contigs with repeats belonging to the significant clusters were also subjected to a BLASTX search against the nr-protein database. Results were first compared with the preliminary taxonomy obtained with CRISPRdb BLAST in order to assess the veracity of these results. See **Supplemental File 5**.

In fact, BLASTX ended up verifying the taxonomy preciously applied to the contigs through their repeat sequences, although not at a species level. The general agreement was made at the genus level. Contigs from clusters E, F and L, which repeat was assigned to *Megamonas hypermegale* ART12/1, had all matches with the genus Megamonas but to *Megamonas funiformis* species. The same happened with cluster K, assigned to the same species within CRISPRdb, but in this case, BLASTX results were more disagreeing, pointing to matches in *Megamonas funiformis* and *Megamonas rupellensis.* Analysing closely the repeat sequences it is possible to see that they are closely related sequence wise, see Figure 11. Repeat sequence from cluster L is exactly the same as the one found and published in CRISPRdb.

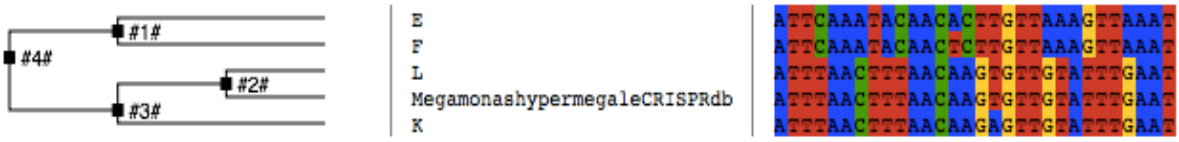


Figure 12 – Repeat alignment with WebPrank (http://www.ebi.ac.uk/goldman-srv/webprank/); the colour code is used to distinguish between nucleotides; Red is for T; Blue is for A; Green is for C and yellow is for G.

An agreement at a species level was found with clusters CB and CQ, previously appointed to *Faecalibacterium prausnitzii* L2-6 and *Escherichia coli* O55:H7, respectively. BLASTX output allowed assigning the majority of contigs to these taxons. Although a taxonomy at a strain level was not always possible.

Concerning cluster BD, associated with *Clostridium saccharobutylicum* DSM 13864, it wasn't possible to reach an acceptable conclusion or agreement since in all the contigs the CRISPR cassette comprised almost the totality of the length of the contig, making impossible the detection or attribution of a taxonomy.

CRISPR-Cas types

A functional CRISPR-*cas* immune system consists of both a CRISPR cassette and *cas* genes (Makarova, Wolf and Koonin 2013). 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 (Makarova, Wolf and Alkhnbashi, et al. 2015). 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 (Makarova, Wolf and Alkhnbashi, et al. 2015). 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.

CRISPRmap

CRISPRs are partially structured non-coding RNAs (ncRNAs), meaning that any type of clustering done regarding the repeats from the CRISPR, should address both the nucleotide sequence and structure. CRISPRmap addresses this, and allows for the "two-way" clustering, with the sequence families and structural motifs.

CRISPR arrays can be themselves classified into 18 structural families and 24 sequence families, divided over 6 Superclasses (A-F), including unclassified repeats. Both structural motifs and sequence families show significant preferential association with particular types or subtypes of *cas* loci, although, different associations might also occur (Lange, et al. 2013).

All unique repeat sequences were analysed with the CRISPRmap, even considering the parameters constraints of the software: the inputted repeat length has to be shorter than 50 bp, which is acceptable for our data, since the longest repeat is 37 bp. For all the detailed results see **Supplemental File 7**.

The principal output of the software is the CRISPRmap tree that contains all the consensus CRISPR sequences in the database clustered together with the input sequences, marked with red lines in the tree. See **Supplemental File 7**.

Among all the sequences analysed, 41 repeats matched repeats already existent in the CRISPRmap database, and 321 sequences were identified has novel repeats. Regarding the latter, to 169 repeats was not recognized a superclass, family or motif. Note that CRISPRmap requires a sequence to match 100 % to a database member in order for it to be identified. Therefore, assigning the feature "novel" to a sequence does not necessarily mean that it is unknown or has no corresponding database sequences. This, however, is not problematic for a correct classification, since the sequence should be classified into the correct cluster nevertheless.

To all assigned repeats, only to 15 wasn't assigned a Superclass. Representatives of all Superclasses were found. The most signified Superclass was Superclass C (with 61 repeats – 34% of all matches), followed by Superclass A (with 43 repeats – 24%) and Superclass F (with 34 repeats – 19%). The least represented class was Superclass D with only 2 linked repeats (1%).

From the 24 conserved repeat sequence families existent, only 7 appeared to be assigned to our set of repeats. The represented families included family F1, F2, F5, F6, F7, F9 and F23, being the most present, sequence family 5 with 14 matches (36% of total assigned repeats). F9 and F23 only had one representative each. It should be noted that only 39 repeats, from a total of 193, were assigned to a sequence family. All repeat sequences with an attributed family were sequences matched in the CRISPRmap database.

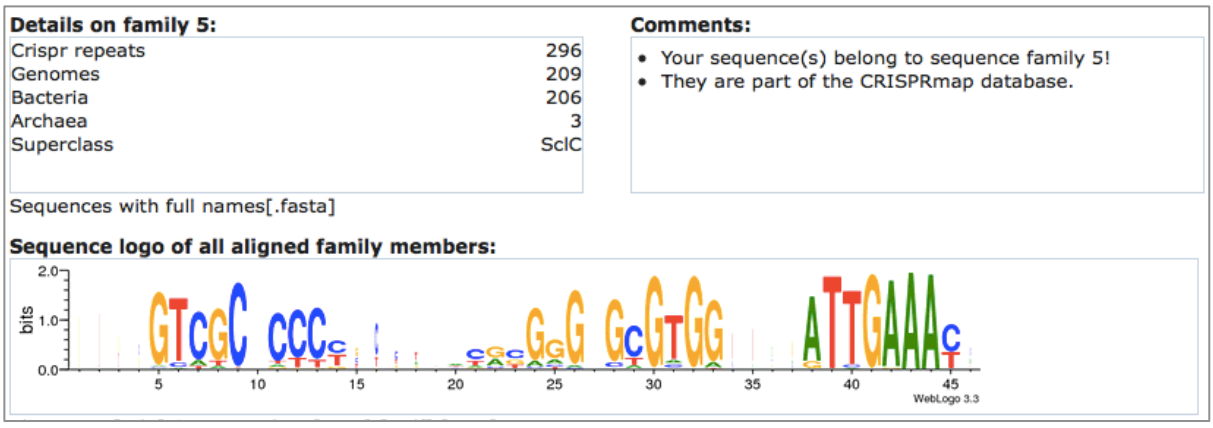


Figure 13 – Details on family 5 provided by the output from CRISPRmap; WebLogo tool is used to generate sequence logos of all aligned family members. Sequence logos present a graphical representation of a multiple sequence alignment of family 5 member repeat sequences, where nucleotides letters that stand out represent the most conserved positions.

Regarding structural motifs, from the existent 18 motifs, only 8 were represented: motif M1, M2, M3, M4, M6, M7, M11 and M16. To more than half of the repeat sequences (about 65%) was assigned the structural motif M3.

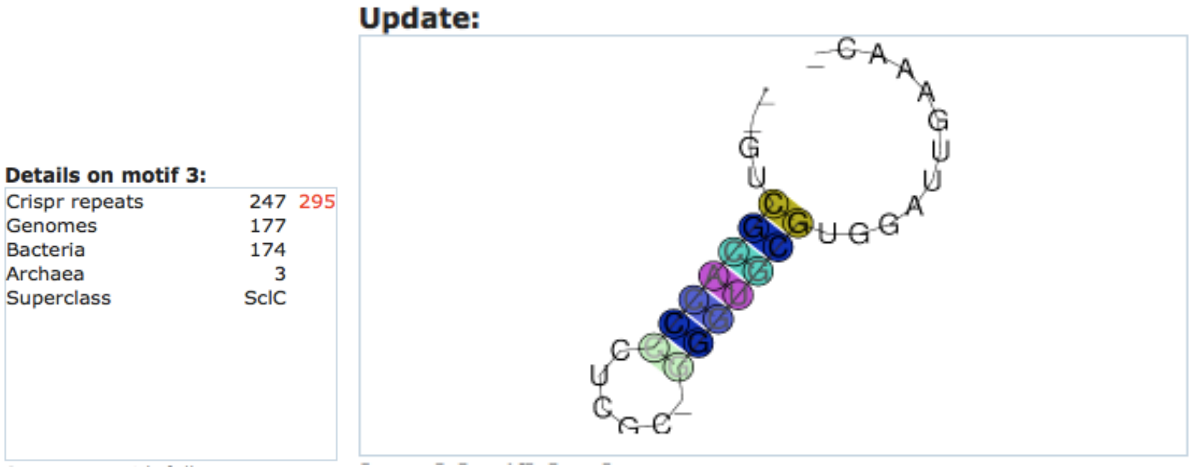


Figure 14 – Details on motif 3 provided by the output from CRISPRmap; Note that the motif was updated after the input of our new set of sequences, not present in their database (48 new repeat sequences); It is possible to observe the palindromic characteristic of CRISPR repeats in the hairpin-like structure of the motif.

In the software, every time a new repeat sequence was attributed to a certain structural motif a multiple sequence alignment (MSA) of the structure motif, as well as the consensus minimum free energy (MFE) structure, both calculated by LocARNA, was given. LocARNA (Will, Joshi, et al. 2012) is a high performance tool designed for multiple alignment of RNA molecules, only using as an input RNA sequences, simultaneously folds and aligns the input sequences. If a new repeat was assigned to the described structure motif and this changed the original structure motif, then the output shows

both the old and new structure motifs (with and without the input repeat sequence(s)). For all changes see **Supplemental File 7**.

Superclasses are constructed and divided according to both sequence and structure similarities. The most represented, Superclass C, contained, as expected, a sequence family, F5, which corresponded to only one structural motif, M3. Only to 8 repeats in this Superclass wasn't attributed a structure motif. On the other "end of the spectrum" is Superclass F. Divergent in sequence and structure conservation, only containing, in our data, one assigned family, F23, and one assigned motif, M4.

Analysing the results as a whole, it is possible to observe that structural motifs were mostly unique for each Superclass and sequence family. The only discernible exception was family F1, which included motifs M1 and M7.

CRISPR-Cas subtypes

The repeat is the central regulatory element in the CRISPR-Cas system, as it serves as a binding template for Cas proteins in all three phases of immunity. The sequence itself and a small hairpin structure, essential in several systems for Cas endoribonucleases binding, are key variables (Brouns, et al. 2008) (Gesner, et al. 2011). Making use of this information, CRISPRmap is also designed for an automated Cas subtype annotation. The classification used for the annotation concerns Makarova et al. classification (Makarova, Haft, et al. 2011).

To the 41 repeat sequences matched to the CRISPRmap database, only 27 had an assigned system type. Among the types found within the set of repeats, the most assigned was type I with 23 matches. Inside type I, subtype I-C stood out with 12 assigned repeats. Type II had one representative (related to subtype II-A), and type III had 3.

Comparing the assigned CRISPR-Cas system types by CRISPRmap with the ones assigned using the BLASTX tool, it is possible to observe a certain kind of concordance between type attributions. Interestingly enough, the unique repeats with subtypes associated by CRISPRmap correspondent generally to unique repeats to who wasn't possible to find a match with BLASTX. To the ones that had matches with both programs, only one didn't find concordance.

In general, also the taxonomy of the repeats present in the CRISPRmap database finds an overlap with our results. For a complete overview of the comparison see **Supplemental File 7**.

Identical CRISPR arrays

The analysis of the obtained CRISPR cassettes also resulted in the finding of three cassettes shared each between two different individuals. These cassettes share both the repeat sequence and the spacers' sequence. On the outset, it should be expected that these CRISPR cassettes are associated with the same bacterial species or strain, since they are present in different contigs from different metagenomes. If that wasn't the case, the acquiring of part or even a complete cassette by a given bacteria could be explained by the occurrence of horizontal transfer (Chakraborty, et al. 2010).

Querying the CRISPR-containing contig sequences with BLASTX, it is possible to confirm that 'scaffold43595_4' and 'scaffold19162_3', from individuals' DLM014 and DLM028, respectively, share the same taxonomy at a species level, being assigned to the bacteria *Prevotella* sp. *oral taxon* 472 str. F0295, from the Bacteroidetes/Chlorobi group order, proving the initial hypothesis.

Contrariwise, the same taxonomy overlap didn't occur with the other two pairs. Contigs 'scaffold31_2' and 'C721646_1', from individuals' DLM014 and DLM028, respectively, only had a match at a genus level. It wasn't possible to compare contigs 'scaffold14960_3' and 'scaffold14702_2' taxonomy (belonging to individuals' DLM021 and DLM028) since the latter had no significant similarities to the nr-protein database from NCBI. To the first, however, was attributed a taxonomy at a genus level to Clostridium.

Reconstructing a CRISPR 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.

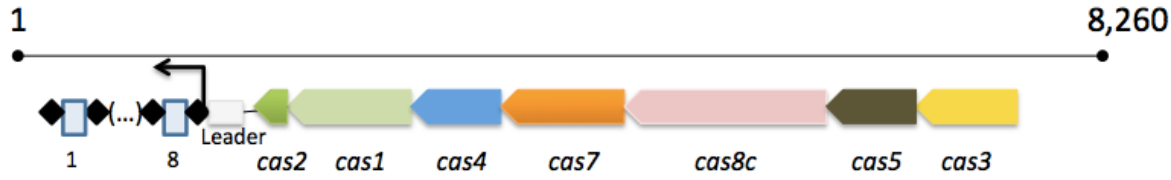
For this purpose we choose four contigs where it was identified a CRISPR cassette and all associated Cas genes. These contigs contain systems representative of type I, type II and putative type V.

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'*) (Makarova, Wolf and Alkhnbashi, et al. 2015). 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*) (Makarova and Koonin, Annotation and classification of CRISPR-Cas systems. 2015). 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, ATTTCAATCCACGTCCCCCGTGCGGGGGACGAC (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..*

A



B

Range: 320 - 890

	REPEAT	SPACER
320	ATTTCAATCCACGTCCCCCGTGCGGGGGACGAC	CGATATGTCCCTATCATAGGCAATAAGTTTGAAA
387	ATTTCAATCCACGTCCCCCGTGCGGGGGACGAC	GCAGACAATATACTCACGCACATCAAGCGCGTAAA
455	ATTTCAATCCACGTCCCCCGTGCGGGGGACGAC	AATGCTAATATATTTGGTGTGTTCTACACTACA
521	ATTTCAATCCACGTCCCCCGTGCGGGGGACGAC	TGGTATCGCCTATCCATCCGAGCATTACCGAGTT
588	ATTTCAATCCACGTCCCCCGTGCGGGGGACGAC	TATATCTGAGCTTACCCCCTGGCAGATCAGGTC
654	ATTTCAATCCACGTCCCCCGTGCGGGGGACGAC	AAGGCGTATATTGAGTTTGCATGCAACATGTTGTT
722	ATTTCAATCCACGTCCCCCGTGCGGGGGACGAC	AATTTTACTGCTTTTGCCTTTTCGTCGTAATACTC
790	ATTTCAATCCACGTCCCCCGTGCGGGGGACGAC	GTTTTTCGGAGACTGTTGCGGACTTAATGGAGAG
857	ATTTCAATCCACGTCCCCCGTGCGGGGGACGAC	
890		

Figure 15 - 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.

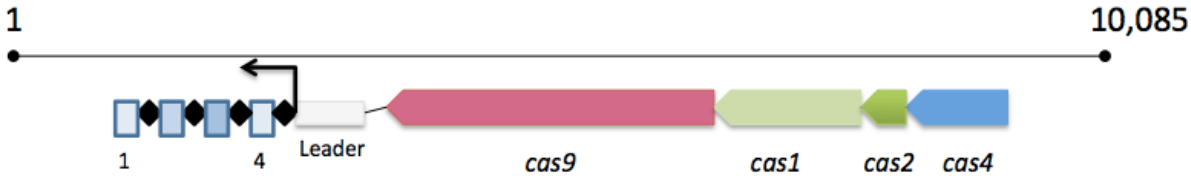
Type II

A CRISPR-Cas system belonging to type II has always the signature gene *cas9.* Two or three core genes, in charge of the adaptation stage, accompany Cas9. The set of core genes present is what differentiates from subtype II-A, II-C and II-B.

In individual DOM16, to which belongs contig 'scaffold7464_7', was discovered a representative of a type II-B system. The latter is composed by the characteristic Cas9, ranging from position 2367 to 6650, and its accompanying core genes: cas1 (6650 to 7639bp); cas2 (7646 to 7930bp) and cas4 (7930 to 8516bp). The corresponding CRISPR array is located upstream and is composed of a set of 5 repeats, GCTTCAATCGTACATTCTTTCAATAACAACTGAAAC, with length 36bp, and 4 spacer sequences. The CRISPR array is harboured by a bacterium belonging to Proteobacteria *phylum*,

which, according to the top hits resulting form the BLASTX search, corresponds to *Parasuterella excrementihominis.*

A



B

Range: 1436 - 1758

	REPEAT	SPACER
1436	GCTTCAATCGTACATTCTTTCAATAACAACTGAAAC	GTGATTCAACTGATTTTTTAATTCCGTCTCTTTCC
1507	GCTTCAATCGTACATTCTTTCAATAACAACTGAAAC	ATTATCACGCTATTATCTTTAACCTACCTCCTCTCG
1579	GCTTCAATCGTACATTCTTTCAATAACAACTGAAAC	TCTTGGACGCTTTCTTAAGACGTTCTAATAATGGGT
1651	GCTTCAATCGTACATTCTTTCAATAACAACTGAAAC	GTTCCATCTAAATCTAATTCTTCAAGATTTTCGAG
1722	GCTTCAATCGTACATTCTTTCAATAACAACTGAAAC	
1758		

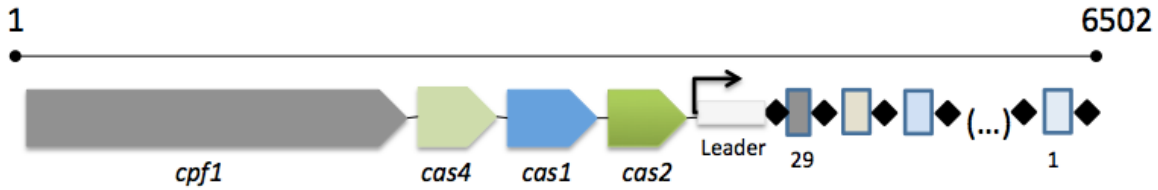
Figure 16 - Schematics of the architecture of an identified type II CRISPR-Cas system, from contig 'scaffold7464_7, originated in individual DOM016; Contig has a total size of 10,085 bp; (**A**) The annotated cas genes, downstream of the array, are constituted by *cas9* gene, characteristic of type II systems, followed by the core cas genes, cas1, cas2 and cas4; (**B**) the CRISPR array, with length 322bp, is composed of 4 spacers (blue coloured rectangles) and 5 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 'oldest' one, and spacer 4 is the newest (it was the last one to be inserted into the array); Note that the scheme is made on scale and does not represent the exact size of the genes or CRISPR array.

Type V

A type V CRISPR-Cas system is immediately identified by the Cas gene *cpf1,* characteristic of this system type. Since it's a relatively new defined type, most of this gene, apart from the RuvC-like domain, is functionally uncharacterized, although it is suggested that it is a functional analogue of Cas9. Putative type V combines Cpf1 (the interference module) with an adaptor module composed by the core genes cas4, cas1 and cas2 (Makarova, Wolf and Alkhnbashi, et al. 2015).

In individual's DLM001, contig 'scaffold23627_1', with a length of 6502 bp, was identified a putative type V system. The latter is composed by the large protein Cpf1, ranging from position 6 to 2675, followed by Cas4 (2695 to 2874bp), Cas1 (2919-3242) and Cas2 (3255 to 3467bp). The CRISPR array is formed by a set of 30 repeats, GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT, with 36bp, and 29 spacers. The CRISPR array is harboured by a bacterium from the Bacteroides *phylum,* which, according to the top hits on BLASTX, corresponds to *Prevotella bryantii*.

A



B

Range: 3698 - 5859

	REPEAT	SPACER
3698	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	GTCTTTTTTTTTCCATTCATTGTATCGTG
3763	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	ACATGGCATCCGAAATTTTGCCTTTGATG
3828	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	ATGGAACTATCTTATATCCTCAATGGGAT
3893	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	GTGTTTAATTTTAATGTCAAGTTATATGG
3958	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	TCCAAGTAGCTGAATCAGTCGGTAAAATA
4023	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	AGAATGGCAAGCTTTATGGTAATAGCAAA
4088	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	GCAAGATAAGCCTCTGTAGCAGCCTCCTCA
4154	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	GATAAGTGCTTTGAGTCTTGTCGTAGGG
4218	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	GTAATATACAGGCTGCTCACCTGCGGCTAG
4284	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	TACTAAGTTGTGTACTTTTGGTTTACAGAT
4350	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	AGATGTATTTCGTGAAGACGACATTATC
4414	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	CTGTGCGAAGGTAGCTTCT
4544	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	CGGAAAATATTTGCACGGAACAAATAAAA
4609	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	TCCTTGCGAAGAATACCACAGAGGTTTTC
4674	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	TCTACACAGCAGGGAGGAATGAATAACAA
4737	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	TCGATGATTCTCCCCTGCTTCGTCAAC
4799	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	GCCATAACTCAAAGATATATGTGGGC
4865	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	TATTCCATAAAGGTTCGATTAAATCAATAG
4929	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	TAATGATGGTGTTTCAATTCATTCTGAT
4995	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	ATTAGCTGCATTGTTTTTGTTACACGTACC
5060	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	GCTATGAAGGTTACAGAGGTGCAGCCACT
5125	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	GTTGAAGAGAGCCTTTTGAAAAAGGGGAA
5189	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	AGATCCATCTTGGGTAACATCATCAATA
5255	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	AGGCGCACCACGATGTTCCCTGTGTCGGAT
5320	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	CGGAGCAAGCCGCATGATGGGTCTCGACG
5385	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	ATCTGACAGTTCGGAAAGACGAGAGTCGT
5477	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	GGCGCATCCCTCGCCAAGGACCTGAA
5513	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	TTGATATTGGCGGACTTGCACAATGTGGCG
5588	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	AAGCTGAACGTTTTAGCTGTATCATTTTT
5655	GGCTAGTAAGCTCTAATAATTTCTACTATTGTAGAT	
5859		

Figure 17 - Schematics of the architecture of an identified type V putative CRISPR-Cas system, from contig 'scaffold23627_1', originated in individual DLM001; Contig has a total size of 6502 bp; (**A**) The annotated cas genes, upstream of the array, are constituted by *cpf1* gene, characteristic of type V systems, followed by the core cas genes, cas4, cas1 and cas2; (**B**) the CRISPR array, with length 2161bp, is composed of 29 spacers (coloured rectangles) and 30 repeat sequences (◆), with an adjacent leader sequence upstream of the first repeat; Spacer 1 is the one furthest from the leader sequence so it is the 'oldest' one, and spacer 29 is the newest; Note that the scheme is made on scale and does not represent the exact size of the genes or CRISPR array

Identification and analysis of protospacers

Taxonomy of protospacer origin and compatibility with the CRISPR-cassette taxonomy

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 (Casey, et al. 2014). 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 (Ogilvie, et al. 2012). 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.*

cRassphage

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 (Dutilh, et al. 2014). 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 (Dutilh, et al. 2014). 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.

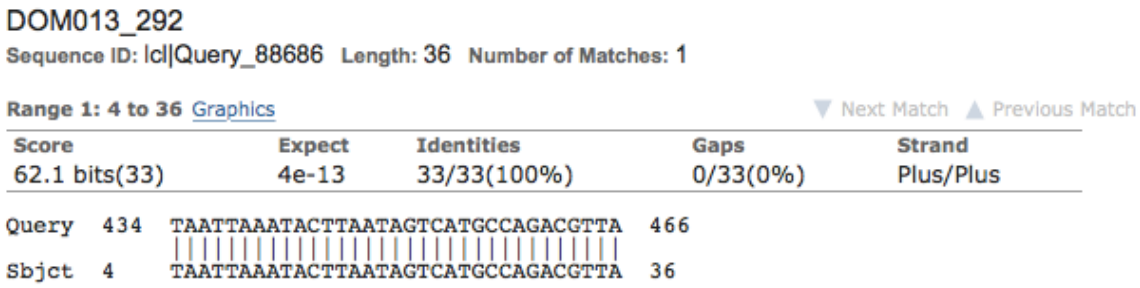


Figure 18 - The output from BLASTN algorithm after 'blasting' our entire set of non-redundant spacers against the 97 kbp genome of bacteriophage crAssphage; 'DOM013_292' identifies the spacer has being from individual sample DOM13.

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 (Rohwer 2003).

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.

Similarity of the spacer composition in the diabetic human gut individual microbiomes

The shared spacers appeared mostly within the same CRISPR array, or in different arrays from different individuals, possibly indicating they were targeted by the same phage or mobile element. It is an interesting occurrence since it was expected for the repetitions to occur more within the same individual's CRISPR arrays, displaying a common phage invasion history.

From all the repeated spacers, 45 (considering only the unique sequence) are shared within the same CRISPR cassette.

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.

Repeat Set

In 2014, Gogleva and colleagues 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.

Table 3 - Characteristics of the available metagenomic datasets subjected to analysis for CRISPR-Cas content; * this dataset refers to the MetaHIT project, and not the HMP, as it is possible to check from the references existent in **(Gogleva, Gelfand and Artamonova 2014)**.

Metagenomic dataset	Number of contigs	Total length	Source	Individuals involved	Sequencing platform	Assembly algorithm
Chinese (**T2D**)	8,039,994 contigs	16,345 Mb	Fecal samples	145 Chinese Han of various ages (14-59)	Illumina GA	SOAPdenovo
The Human Microbiome Project (**HMP**)*	1,889,651 contigs	3,732 Mb	Fecal samples	124 Europeans of various ages (18-69)	Illumina GA	MetaMos
Healthy Human Gut Metagenomes (**JPN**)	353,805 contigs	463 Mb	Fecal samples	13 Japanese individuals (6 months – 45 years)	MegaBACE4500 sequencer (GE Healthcare)	PCAP
Distal Gut metagenomic project (**DG**)	22,508 contigs	336 Mb	Fecal samples	2 healthy adults	ABI 3730x1 DNA analyzer	Celera Assembler

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 (Li, et al. 2012). 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) (J. Qin, Y. Li, et al. 2012).

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). 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 (J. Qin, R. Li, et al. 2010) 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.

Spacers 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 (Stern, Mick, et al. 2012). 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.

General Conclusion 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.

References

Ackermann, H. W. “Bacteriophage taxonomy.” *Microbiology Australia*, 2011: 90-94.

Altschul, S. F., Madden, T. L., Schäffer, 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 Research* 25 (1997): 3389-402.

Arslan, Z., Wurm, R., Brener, O., Ellinger, P., Nagel-Steger, L., Oesterhelt, F., Schmitt, L., Willbold, D., Wagner, R., Gohlke, H., Smits, S. H. and Pul, U. “Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2.” *Nucleic Acids Research* 41, no. 12 (2013): 6347-59.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., Horvath, P. “CRISPR provides acquired resistance against viruses in prokaryotes.” *Science* 315 (2007): 1709-12.

Barrangou, R. “The Roles of CRISPR-Cas systems in adaptive immunity and beyond.” *Current Opinion in Immunology* 32 (2015): 36-41.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W. “GenBank.” *Nucleic Acids Research* 41, no. (Database issue) (2013): D36–D42.

Berg Miller, M. E., Yeoman, C. J., Chia, N., Tringe, S. G., Angly, F. E., Edwards, R. A., Flint, H. J., Lamed, R., Bayer, E. A., White, B. A. “Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome.” *Environmental Microbiology* 14, no. 1 (2012): 207-27.

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

Bland , C., Ramsey , T. L., Sabree , F., Lowe , M., Brown, K., Kyrpides, N. C., Hugenholtz, P. “CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats.” *BMC Bioinformatics* 8, no. 1 (2007): 209.

BLAST CRISPR. http://crispr.u-psud.fr/crispr/BLAST/CRISPRsBlast.php.

BLASTClust tool. http://www.animalgenome.org/blast/doc/blastclust.html.

Bolduc, B., Shaughnessy, D. P., Wolf, Y. I., Koonin, E. V., Roberto, F. F. and M. Young. “Identification of novel positive-strand RNA viruses in metagenomic analysis of archaea-dominated Yellowstone hot springs.” *Journal of Virology* 86, no. 10 (2012): 5562-73.

Bolotin, A., Quinquis,B., Sorokin, A. and S. D. Ehrlich. “Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin.” *Microbiology* 151 (2005): 2551-61.

Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. and A. R. Davidson. “Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system.” *Nature* 493, no. 7432 (2013): 429-32.

Botsaris, G., Liapi, M., Kakogiannis, C., Dodd, C. E. and C. E. Rees. “Detection of Mycobacterium avium subsp. paratuberculosis in bulk tank milk by combined phage-PCR assay: evidence that plaque number is a good predictor of MAP.” *International Journal of Food Microbiology* 164, no. 1 (2013): 76-80.

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P. L., Dickman, M. J., Makarova, K. S., Koonin, E. V., van der Oost, J. “Small CRISPR RNAs guide antiviral defense in prokaryotes.” *Science* 321 (2008): 960-64.

Candela, M., Biagi,E., Maccaferri, S., Turroni, S. and P. Brigidi. “Intestinal microbiota is a plastic factor responding to environmental changes.” *Trends in Microbiology* 20 (2012): 385–91.

Cantarel, B. L., Lombard, V. and B. Henrissat. “Complex Carbohydrate Utilization by the Healthy Human Microbiome.” *PLoS ONE* 7, no. 6 (2012): e28742.

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.

Chakraborty, S., Snijders, A. P., Chakravorty, R., Ahmed, M., Tarek, A. M. and M. A. Hossain. “Comparative network clustering of direct repeats (DRs) and cas genes confirms the possibility of the horizontal transfer of CRISPR locus among bacteria.” *Molecular phylogenetics and evolution* 56, no. 3 (2010): 878-87.

Chibani-Chennoufi, S., Bruttin, A., Dillmann, M. L. and H. Brüssow. “Phage-Host Interaction: an Ecological Perspective.” *Journal of Bacteriology* 186, no. 12 (2004): 3677–86.

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.

Clokie, M. R. J., Millard, A. D., Letarov, A. V. and S. Heaphy. “Phages in nature.” *Bacteriophage* 1, no. 1 (2011): 31-45.

Comas, I., Homolka, S., Niemann, S. and S. Gagneaux. “Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in Mycobacterium tuberculosis Highlights the Limitations of Current Methodologies.” *PLoS ONE* 4, no. 11 (2009): e7815.

CRISPRtionary. http://crispr.u-psud.fr/CRISPRcompar/Dict/Dict.php.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J., Charpentier, E. “CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III.” *Nature* 471, no. 7340 (2011): 602-607.

Deveau, H. , Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P., Moineau, S. “Phage response to CRISPR-encoded resistance in Streptococcus thermophilus.” *Journal of Bacteriology* 190 (2008): 1390–1400.

Dsouza, M., Larsen, N. and R. Overbeek. “Searching for patterns in genomic data.” *Trends in Genetics* 13, no. 12 (1997): 497-8.

Durand, P., Mahe, F., Valin, A. S. and J. Nicolas. “Browsing repeats in genomes: Pygram and an application to non-coding region analysis.” *BMC Bioinformatics* 7 (2006): 477.

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., Barr, J. J., Speth, D. R., Seguritan, V., Aziz, R. K., Felts, B., Dinsdale, E. A., Mokili, J. L., Edwards, R. A. “A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes.” *Nature Communications* 5 (2014): 4498.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E., Relman, D. A. “Diversity of the Human Intestinal Microbial Flora.” *Science* 308 (2005): 1635-1638.

Edgar, R. C. “MUSCLE: a multiple sequence alignment method with reduced time and space complexity.” *BMC Bioinformatics* 19, no. 5 (2004): 113.

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

Emerson, J. B., Andrade, K., Thomas, B. C., Norman, A., Allen, E. E., Heidelberg, K. B., Banfield, J. F. “Virus-host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia.” *Archaea* 2013, no. 370871 (2013): 1-12.

Erdmann, S., and R. A. Garrett. “Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms.” *Molecular Microbiology* 85, no. 6 (2012): 1044-56.

FlankAlign. http://crispr.u-psud.fr/crispr/CRISPRFlankingAlignment.php.

Gesner, E. M., Schellenberg, M. J., Garside, E. L., George, M. M. and A. M. Macmillan. “Recognition and maturation of effector RNAs in a CRISPR interference pathway.” *Nature Structural & Molecular Biology* 18 (2011): 688-92.

Godde, J. S., and A. Bickerton. “The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes.” *Journal of Molecular Evolution* 62, no. 6 (2006): 718-29.

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

Grissa, I., Vergnaud, G. 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.

Guinane, C.M. and P. D. Cotter. “Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ.” *Therapeutics Advances in Gastroenterology* 6, no. 4 (2013): 295-308.

Haft, D. H., Selengut, J., Mongodin, E. F. and K. E. Nelson. “A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.” *PLoS Computational Biology* 1 (2005): e60.

Hale, C. R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A. M., Glover III, C. V. C., Graveley, B. R., Terns, R. M., Terns, M. P. “Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs.” *Molecular Cell* 45, no. 3 (2012): 292-302.

Hatoum-Aslan, A., Maniv, I., Samai, P. and L. A. Marraffini. “Genetic Characterization of Antiplasmid Immunity through a Type III-A CRISPR-Cas System.” *Journal of Bacteriology* 196, no. 2 (2014): 310-17.

Hatoum-Aslan, A., K. L. Palmer, M. S. Gilmore, and L. A. Marraffini. “Type III CRISPR-Cas Systems and the Roles of CRISPR-Cas in Bacterial Virulence.” In *CRISPR-Cas Systems.*, by R. Barrangou and J. van der Oost, 201-19. 2013.

Heler, R., Samai, P., Modell, J. W., Weiner, C., Goldberg, G. W., Bikard, D. and L. A. Marraffini “Cas9 specifies functional viral targets during CRISPR-Cas adaptation.” *Nature* 519, no. 7542 (2015).

Hermans, P. W., van Soolingen, D., Bik, E. M., de Haas, P. E., Dale, J. W., van Embden, J. D., “Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains.” *Infection and Immunity* 59 (1991): 2695-705.

Hofacker, I. L., and P. F. Stadler. “Memory efficient folding algorithms for circular RNA secondary struc- tures.” *Bioinformatics* 22, no. 10 (2006): 1172-6.

Hoffman, A. R., Proctor, L. M., Surette, M. G. and J. S. Suchodolski. “The Microbiome: The Trillions of Microorganisms That Maintain Health and Cause Disease in Humans and Companion Animals.” *Veterinary Pathology*, 2015.

Hollister, E. B., Brooks, J. P. and T. J. Gentry. “Nucleic Acid-Based Methods of Analysis.” In *Environmental Microbiology*, by I. L. Pepper, C. P. Gerba and T. J. Gentry, 271-305. Elsevier, 2015.

Horvath, P., and R. Barrangou. “CRISPR/Cas, the immune system of bacteria and archaea.” *Science* 327 (2010): 167-170.

Hyman, P. and S. T. Abedon. *Bacteriophage Host Range ﷺContents and Bacterial Resistance.* Vol. 70, in *Advances in Applied Microbiology*, by A. I. Laskin and G. M. Gadd, 217-48. Elsevier Science, 2010.

Ishino, Y., Shinagawa,H., Makino, K., Amemura, M. and A. Nakata. “Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product.” *Journal of Bacteriology* 169 (1987): 5429-33.

Jansen, R., van Embden, J. D., Gaastra, W. and L. M. Schouls. “Identification of genes that are associated with DNA repeats in prokaryotes.” *Molecular Microbiology* 43, no. 6 (2002): 1565-75.

Jassim, S. A. A. and R. G. Limoges. “Natural solution to antibiotic resistance: bacteriophages ‘The Living Drugs’.” *World Journal of Microbiology and Biotechnology* 30, no. 8 (2014): 2153–70.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M. and J. Doudna . “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.” *Science* 337, no. 6096 (2012): 816–821.

Jore, M. M., Brouns, S. J. J. and J. van der Oost. “RNA in Defense: CRISPRs Protect Prokaryotes against Mobile Genetic Elements.” *Cold Spring Harbor Perspectives in Biology* 4, no. 6 (2012): a003657.

Kelly, D., and I. E. Mulder. *Microbiome and immunological interactions* (Nature Reviews) Suppl. 1 (2012): S18-30.

Koskella, B., and M. A. Brockhurst. “Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities.” *FEMS Microbiology Reviews* 38, no. 5 (2014): 916–31.

Krupovic, M., Prangishvili, D., Hendrix, R. W. and D. H. Bamford. “Genomics of Bacterial and Archaeal Viruses: Dynamics within the Prokaryotic Virosphere.” *Microbiology and Molecular Biology Reviews* 75, no. 4 (2011): 610-35.

Kunin, V., Sorek, R. and P. Hugenholtz. “Evolutionary conservation of sequence and secondary structures in CRISPR repeats.” *Genome Biology* 8, no. 4 (2007): R61.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich R. “REPuter: Fast Computation of Maximal Repeats in Complete Genomes.” *Bioinformatics* 15, no. 5 (1999): 426-7.

Labrie, S. J., Samson, J. E. and S. Moineau. “Bacteriophage resistance mechanisms.” *Nature Reviews Microbiology* 8 (2010): 317-27.

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

LeBlanc, J. G., Milani, C., de Giori, G. S., Sesma, F., van Sinderen, D., Ventura, M. “Bacteria as vitamin suppliers to their host: a gut microbiota perspective.” *Current Opinion in Biotechnology* 24 (2013).

Lederberg, J. “Infectious history.” *Science* 288 (2000): 287-293.

Ley, R. E., Peterson, D. A. and J. I. Gordon. “Ecological and evolutionary forces shaping microbial diversity in the human intestine.” *Cell* 124 (2006): 837-848.

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

Li, S., Guan, Y., Zhang, W., Zhang, F., Cai, Z., Wu, W., Zhang, D., Jie, Z., Liang, S., Shen, D., Qin, Y., Xu, R., Wang, M., Gong, M., Yu, J., Zhang, Y., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Zhang, M., Zhang, Z., Chen, H., Qin, J., Li, Y., Wang, J. ”Type 2 diabetes gut metagenome (microbiome) data from 368 Chinese samples.” (2012) http://gigadb.org/dataset/100036.

Louwen, R., Staals, R. H. J., Endtz, H. P., van Baarlen, P. and J. van der Oost. “The Role of CRISPR-Cas Systems in Virulence of Pathogenic Bacteria.” *Microbiology and Molecular Biology Reviews* 78, no. 1 (2014): 74-88.

Makarova, K. S. and E. V. Koonin. “Annotation and classification of CRISPR-Cas systems.” *Methods in Molecular Biology* 1311 (2015): 47-75.

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Terns, R. M., White, M. F., Yakunin, A. F., Garrett, R. A., van der Oost, J., Backofen, R., Koonin, E. V. “An updated evolutionary classification of CRISPR-Cas systems.” *Microbiology* 13 (2015): 1-15.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J., Wolf, Y. I., van der Oost, J., Koonin, E. V. “Evolution and classification of the CRISPR–Cas systems.” *Nature Reviews Microbiology* 9 (2011): 467-77.

Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wlf, Y. I. and E. V. Koonin. “A putative RNA-interference-based immune system in prokaryotes: computional analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.” *Biology Direct* 1, no. 7 (2006).

Makarova, K. S., V. Anantharaman, L. Aravind, and E. V. Koonin. “Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes.” *Biology Direct* 14, no. 7 (2012): 40.

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.

Metagenomics: Sequences from the Environment . Bethesda (MD): National Center for Biotechnology Information (US). 2006. http://www.ncbi.nlm.nih.gov/books/NBK6858/.

Metzker , M. L. “Sequencing technologies – the next generation.” *Nature Reviews* 11, no. 1 (2010).

Mojica, F. J., Diez-Villasenor, C., Soria, E. and G. Juez. “Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria.” *Molecular Microbiology* 36 (2000): 244-6.

Mojica, F. J., Diez-Villaseñor, C., Garcia-Martinez, J. and E. Soria. “Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.” *Journal of Molecular Evolution* 60, no. 2 (2005): 174-82.

Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J. and C. Almendros. “Short motif sequences determine the targets of the prokaryotic CRISPR defence system.” *Microbiology* 155 (2009): 733-40.

Mojica, F. J., Ferrer, C., Juez, G. and F. Rodriguez-Valera. “Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning.” *Molecular Microbiology* 17 (1995): 85-93.

NIH HMP Working Group *et al.* “The NIH Human Microbiome Project.” *Genome Research* 9, no. 12 (2009): 2317-23.

Nunez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V., Davies, C. W. and J. A. Doudna. “Cas 1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity.” *Nature Structural & Molecular Biology* 21 (2014): 528–34.

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.

Pedersen, M. B., Jensen, P. R., Janzen, T. and D. Nilsson. “Bacteriophage Resistance of a ΔthyA Mutant of Lactococcus lactis Blocked in DNA Replication.” *Applied and Environmental Microbiology* 68, no. 6 (2002): 3010–23.

Pougach, K. S., Lopatina, A. V. and K. V. Severinov. “CRISPR Adaptive Immunity Systems of Prokaryotes.” *Molecular Biology* 46, no. 2 (2012): 195-203.

Pourcel, C. and C. Drevet. “Occurrence, Diversity of CRISPR-Cas Systems and Genotyping Implications.” In *CRISPR-Cas Systems*, by R. Barrangou and J. van der Oost , 33-59. Springer Berlin Heidelberg, 2013.

Pourcel, C., Salvignol, G. and G. Vergnaud. “CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies.” *Microbiology* 151 (2005): 653-63.

Pride, D. T., Sun, C. L., Salzman, J., Rao, N., Loomer, P., Armitage, G. C., Banfield, J. F., Relman, D. A. “Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time.” *Genome Research* 21, no. 1 (2011): 126-36.

Qin, J., Li, R., Raes, J., MetaHIT Consortium; Bork, P., Ehrlich, S. D., Wang, J. “A human gut microbial gene catalogue established by metagenomic sequencing.” *Nature* 464, no. 7285 (2010): 59-65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., *et al.* “A metagenome-wide association study of gut microbiota in type 2 diabetes.” *Nature* 490 (2012): 55-60.

Quercia, S., Candela, M., Giuliani , C., Turroni, S., Luiselli, D., Rampelli, S., Brigidi, P., Franceschi , C., Bacalini, M. G., Garagnani, P., Pirazzini, C. “From lifetime to evolution: timescales of human gut microbiota adaptation.” *Frontiers in Microbiology* 5 (2014): 587.

Rappé, M. S. and S. J. Giovannoni . “The uncultured microbial majority.” *Annual Reviews in Microbiology* 57 (2003): 369-94.

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.

Richter, C., Chang, J. T. and P. C. Fineran. “Function and Regulation of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) / CRISPR Associated (Cas) Systems.” *Viruses* 4, no. 10 (2012): 2291-311.

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): 1710–24.

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

Round, J. L. and S. K. Mazmanian. “The gut microbiome shapes intestinal immune responses during health and disease.” *Nature Reviews Immunology* 9, no. 5 (2009): 313–323.

Rusch , D. B. *et al*. “The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.” *PLoS Biology* 5, no. 3 (2007): e77.

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.

Sabat, A. J., Budimir, A., Nashev, D., Sá-Leão, R., van Dijl, J. M., Laurent, F., Grundmann, H., Friedrich, A. W., ESCMID, Study Group of Epidemiological Markers (ESGEM) “Overview of molecular typing methods for outbreak detection and epidemiological surveillance.” *Euro Surveillance* 18, no. 4 (2013): 20380.

Salazar, N., Arboleya, S., Valdés, L., Stanton, C., Ross, P., Ruiz, L., Gueimonde, M., de los Reyes-Gavilán, C. G. “he human intestinal microbiome at extreme ages of life. Dietary intervention as a way to counteract alterations.” *Frontiers in Genetics* 5 (2014): 1-9.

Samson, J. E., A. H. Magadán, M. Sabri, and S. Moineau. “Revenge of the phages: defeating bacterial defences.” *Nature Reviews Microbiology* 11 (2013): 1-13.

Semenova, E., Matthijs, M. J., Jore, M., Datsenko, K. A., Semenova, A., Westra, E. R., Wanner, B., van der Oost, J., Brouns, S. J. J., Severinov, K. “Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence.” *Procedings of the National Academy of Sciences of the U S A* 108, no. 25 (2011): 10098–10103.

Shariat, N. and E. G. Dudley. “CRISPRs: Molecular Signatures Used for Pathogen Subtyping.” *Applied Environmental Microbiology* 80, no. 2 (2014): 430-39.

Sikunas, T., Gasiunas, G., Waghmare, S. P., Dickman, M. J., Barrangou, R., Horvath, P., Siksnys, V. “In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus.” *EMBO Journal* 32 (2013): 385-94.

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

Sorokin, V. A., Gelfand, M. S. and I. I. Artamonova. “Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome.” *Applied and Environmental Microbiology* 76, no. 7 (2010): 2136-44.

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 (2012): 1985-94.

Stern, A., Keren, L., Wurtzel, O., Amitai, G. and R. Sorek. “Self-targeting by CRISPR: gene regulation or autoimmunity?” *Trens in Genetics* 26 (2010): 335-40.

Streicher, E. M., Victor, T. C., van der Spuy, G., Sola, C., Rastogi, N., van Helden, P. D., Warren, R. M. “Spoligotype signatures in the Mycobacterium tuberculosis complex.” *Journal of Clinical Microbiology* 45, no. 1 (2007): 237-40.

Sturino, J. M. and T. R. Klaenhammer. “Engineered bacteriophage-defence systems in bioprocessing.” *Nature Reviews Microbiology* 4 (2006): 395-404.

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

The NIH HMP Working Group, *et al*. “The NIH Human Microbiome Project.” *Genome Research* 19, no. 12 (2009).

Thompson, J. D., Higgins, D. G. and T. J. Gibson. “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.” *Nucleic Acids Research* 22, no. 22 (1994): 4673-80.

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

Tyson , G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev V. V., Rubin, E. M., Rokhsar, D. S., Banfield, J. F. “Community structure and metabolism through reconstruction of microbial genomes from the environment.” *Nature* 428, no. 6978 (2004): 37-43.

van der Oost, J., M. M. Jore , E. R. Westra , M. Lundgren , and S. J. Brouns. “CRISPR-based adaptive and heritable immunity in prokaryotes.” *Trends in Biochemstry Science* 34 (2009): 401-407.

van Erp, P. B. G., Bloomer, G., Wilkinson, R. and B. Wiedenheft. “The history and market impact of CRISPR RNA-guided nucleases.” 12 (2015).

Vmatch. http://www.vmatch.de.

Wei, Y., Chesne, M. T., Terns, R. M. and M. P. Terns. “Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus.” *Nucleic Acids Research*, 2015: 1749-1758.

Westra, E. R., Swarts, D. C., Staals, R. H., Jore, M. M., Brouns, S. J. and J. van der Oost. “The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity.” *Annual Reviews in Genetics* 46 (2012): 311-39.

Wiedenheft, B., Sternberg, S. H. and J. A. Doudna. “RNA-guided genetic silencing systems in bacteria and archaea.” *Nature* 402 (2012): 331-8.

Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F. and R. Backofen. “Inferring Non-Coding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering.” *PLoS Computational Biology* 3, no. 4 (2007): e65.

Will, S., Joshi,T., Hofacker, I. L., Stadler, P. F. and R. Backofen. “LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs.” *RNA* 18, no. 5 (2012): 900-14.

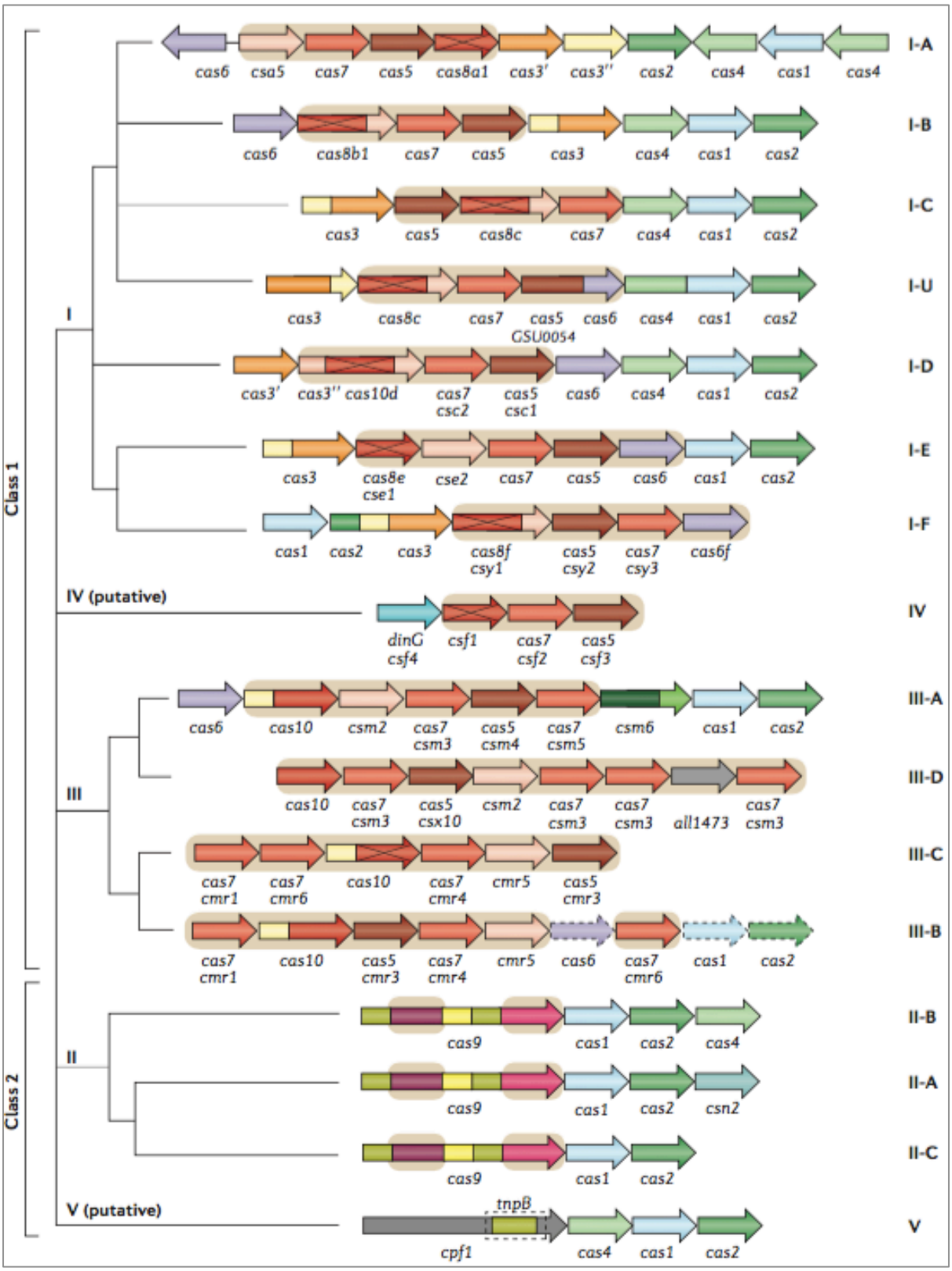
Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., *et al.* “Human gut microbiome viewed across age and geography.” *Nature* 486 (2012): 222–7.

Yosef, I., Goren, M. G. and U. Qimron. “Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli.” *Nucleic Acids Research* 40 (2012): 5569-76.

Supplemental Figures

Supplemental Figure 1

Detailed architecture of all CRISPR-Cas types and subtypes classified in 2015 by Makarova and colleagues, in reference (Makarova, Wolf and Alkhnbashi, et al. 2015).



Supplemental Files

In CD-ROM

Supplemental File 1 (file “Supplemental XCL Table 1- Chinese DNA samples info.xlsx”)

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 (file “Supplemental XCL Table 2 – Collection of predicted reliable CRISPR cassettes.xlsx”)

PILER-CR and CRT predicted CRISPR cassettes.

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: (file “Supplemental XCL Table 3 – Set of non-redundant repeat sequences.xlsx”)

Collection of all repeats and construction of a non-redundant set of CRISPR cassettes repeat sequences.

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: (file “Supplemental XCL Table 4 – Repeat sequence BLAST against CRISPR database, CRISPRdb.xlsx”)

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: (file “Supplemental XCL Table 5 – Set of significant clusters from the set of non-redundant repeats.xlsx”)

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: (file “Supplemental XCL Table 6 – Unique repeats collection and BLASTX output.xlsx”)

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

Unique repeats are shown in more detail, with concern to their sequence, origin contig and contig length, CRISPR length (including start and end positions). There are columns dedicated to the attribution of the CRISPR-Cas type (or subtype), as well as the taxonomy assigned to each individual contig.

Supplemental File 7: (file "Supplemental XCL Table 7 – CRISPRmap output.xlsx")

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.

A link to the detailed output provided from CRISPRmap is provided in sheet (A).

In sheets (C) and (D), some further details are presented for the sequence families and structural motifs, respectively.

Supplemental File 8: (file "Supplemental XCL Table 8 – Collection of Spacers and CRISPRTarget output.xlsx")

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: (file "Supplemental XCL Table 9 – Comparative analysis.xlsx")

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

The results for the comparative analysis conducted for both the non-redundant set of repeats and spacers. Sheet (A) comprises the comparative analysis between the results from Gogleva *et al* (Gogleva, Gelfand and Artamonova 2014) relative to the metagenomic datasets from HMP, JPN and DG; Sheet (B) comprises the comparative analysis relative to the spacer sequences, a comparison between our set of spacers and the set of spacers predicted by Stern *et al*. (Stern, Mick, et al. 2012).