氏　　名	杉本 竜太
学位(専攻分野)	博士(理学)
学 位 記 番 号	総研大乙第 276 号
学位授与の日付	2022 年 9 月 28 日
学位授与の要件	学位規則第6条第2項該当
学 位 論 文 題 目	Development of human endogenous retrovirus database, and comprehensive discovery of CRISPR-targeted mobile genetic elements in the human gut metagenome
論 文 審 査 委 員	主　査　仁木 宏典 遺伝学専攻 教授 中村 保一 遺伝学専攻 教授 工樂 樹洋 遺伝学専攻 教授 森 宙史 遺伝学専攻 准教授 Parrish, Nicholas 理化学研究所 生命医科学研究センター チームリーダー

（様式３）

Summary of Doctoral Thesis

Name in full　Sugimoto, Ryota

Title　Development of human endogenous retrovirus database, and comprehensive discovery of CRISPR-targeted mobile genetic elements in the human gut metagenome

Viruses are the most abundant and diverse genetic entities on Earth. Uncovering such enormous diversity is the key to understanding the origins and evolution of viruses. To collect diverse viral genomes, metagenomics has been utilized in numerous studies. From a given environmental sample, nucleic acids are extracted and sequenced. These sequences presumably include all genetic information of cellular and non-cellular entities existing in the sample. From them, one can computationally filter and extract genetic information of viruses that could not be sequenced otherwise. Such approaches are often referred to as “viral mining”, and have been applied to diverse environmental samples which led to extract tens of thousands of novel viral genomes.

Viral mining is a computational method to extract viral genomic sequences from metagenomes. The viral signature genes such as capsids, RNA-dependent RNA polymerase, and DNA packaging genes are commonly used criteria to detect viral genomes from metagenomic sequences. From a given metagenomic contig, predicted amino acid sequences are searched to reference viral protein databases and if a contig is enriched with the viral signatures genes, one could argue that the contig is positively viral. This implementation has proven highly successful by increasing the known viral genome diversity by orders of magnitudes throughout the multiple studies. However, such a protein homology-based method highly relies on the availability of well-annotated viral protein sequences in the database, and arguably reference viral genomes recorded in the database are highly biased. For example, there are currently 14,073 *Caudovirales* genomes recorded in the NCBI database. In contrast, only 90 *Tubularvirales*, also known as filamentous phages, species genomes are recorded in the

NCBI database. This discrepancy is even notable for RNA bacteriophages. Currently, there are only two RNA bacteriophage families; *Cystoviridae* and *Fiersviridae,* which comprise 27 species are recorded in the NCBI database. Therefore, the viral mining methods that relied on reference genomes might fail to capture entire viral diversity, especially for the least characterized viral lineages. Furthermore, even for the well-characterized viral clade such as *Caudovirales*, it was shown that the detection of viral signature genes such as capsid can be challenging due to their extreme sequence diversity.

Thus, I sought to develop a non-reference-based analytical pipeline to detect viral sequences from metagenomes; however, I was tasked with the question "What information could be used for this purpose?" The underlying biology of the clustered regularly interspaced short palindromic repeats (CRISPR) system, a prokaryotic form of adaptive immunological memory, provides a potential resource in this context. After viral infection or horizontal plasmid transfer, some archaeal and bacterial cells incorporate fragments of "nonself" genetic materials in specialized genomic loci between CRISPR direct repeats (DRs). The incorporated sequences, called "spacers," are identical to part of the previously infecting mobile genetic element. Thus, the genetic information encoded in CRISPR spacers can be inferred as likely viral and distinguishable from the genetic material of the organism encoding CRISPR, which is most often cellular, but potentially also viral.

CRISPR spacers have previously been extracted from assembled bacterial genomes to assess CRISPR "dark matter", revealing that 80%–90% of identifiable material matches known viral genomes. In the current study, I extended this conceptual approach to the enormous amount of unassembled short-read metagenomic data publicly available from the sequence repositories.

By analyzing publicly available human gut metagenome reads and the contigs assembled from them, I extracted 11,391 terminally redundant (TR) CRISPR-targeted

sequences ranging from 894 to 292,414 bases. These sequences are expected to be complete or near-complete circular genomes that can be linked to their CRISPR-targeting hosts. The discovered sequences include 2,154 tailed-phage genomes, together with 257 complete crAssphage genomes, 11 genomes larger than 200 kilobases (kb), 766 Microviridae genomes, 56 Inoviridae genomes, 5,658 plasmid-like genomes, and 2,757 uncharacterized genomes. Although the majority of the discovered sequences that were larger than 20 kb were mostly characterized by viral or plasmid genomes, a substantial portion of sequences smaller than 20 kb was not recorded in either plasmid or viral databases. Furthermore, some previously uncharacterized small genomes had notably low coding ratios, which indicates that these elements might have unknown non-coding genetic features.

Additionally, I investigated CRISPR-targeted RNA sequences in the human gut. Using publicly available human intestine microbe omics data, I extracted non-transcribed RNA sequences by comparing metagenome and metatranscriptome sequences. From them, I searched for CRISPR-targeted RdRP coding sequences. Interestingly, I found some *Picobirnaviridae* species, which host is currently controversial, are being targeted by the CRISPR-Cas system.

These results demonstrate that our pipeline can discover CRISPR-targeted mobile genetic elements (MGEs) either previously characterized or uncharacterized.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏 名　　杉本　竜太

Title
論文題目　　Development of human endogenous retrovirus database, and comprehensive discovery of CRISPR-targeted mobile genetic elements in the human gut metagenome

細菌は常にファージ（細菌のウィルス）との死闘を繰り広げ進化している。毎日、海洋の細菌の約 3 分の 1 はファージによって殺されていると考えられる。細菌はこのファージの脅威に対抗するため、独自の抵抗機構を作り上げた。それが細菌の免疫とも呼ばれる CRISPR-Cas システムである。CRISPR-Cas システムでは、まず細菌は侵入してきたファージのゲノム DNA を切断し、その断片を細菌自身のゲノムに埋め込み、生き残った細菌にその感染の記録を残す。こうして、次に侵入してきた時、いち早くそのファージを感知し増殖を阻止するのである。細菌ゲノムのファージゲノム情報が埋め込まれた領域を CRISPR（クリスパー）と呼ぶ。この CRISPR にはその細菌が過去に感染した様々なファージのゲノムの配列が記憶されている。申請者は CRISPR に書き込まれたこの配列を読み取れば、未知のファージのゲノム情報を検出できると考えた。そこで、多くの細菌のゲノム情報やメタゲノム情報のデータベースから CRISPR 配列を見出し､その中の CRISPR スペーサー配列部分からファージのゲノム配列情報を取り出し、CRISPR スペーサー配列と由来した細菌系統のデータベースを構築した。申請者が構築したデータベースは既知のファージのゲノム情報を必要としないため、これを用いて未だ培養もされていない多数の多様なファージを検出することができる点で非常に有用である。実際 CRISPR スペーサーを利用したこの検出方法で公開済みのメタゲノムデータベースから、既存の方法では発見できなかった新規のファージゲノムを多数検出することに成功している。ヒトの腸内細菌のメタゲノムデータベースを利用した結果、CRISPR 配列から 11,391 のファージゲノム情報を抽出し、そのうち、2.154 は尾部を持つバクテリオファージのゲノム、257 は crAssphage のゲノム、766 は Microviridae ファージのゲノム、56 は Inoviridae ファージのゲノム、95 はこれまでに全く知られていないファージのゲノムをそれぞれ見出している。また 11 のゲノムは 200kbp を超えるものであった。加えて、CRISPR 配列を持つ宿主である細菌のゲノムの情報を手がかりに、本研究で見出したファージのゲノムの約７０％についてその宿主を分類上の門レベルで明らかにしている。

申請者は以上のように公開データベースのバクテリアのメタゲノム情報を独自に大規模情報解析し、ヒト腸内で細菌に感染するファージの多様性について新規性の高い知見を明らかにした。加えて、メタゲノム配列とメタトランスクリプトーム配列のデータを組み合わせて、ヒト腸内の RNA ファージについての解析を行うなど挑戦的な研究にも着手し先駆的な研究にも取り組むなど、博士論文として非常に優れたものと認められる。