博士論文（要約）

Toward complete reconstruction of repetitive sequences in metagenome and centromere

(メタゲノムおよびセントロメアにおける反復配列の
完全な再構成に向けて**)**

鈴木　慶彦

Toward complete reconstruction of repetitive sequences in metagenome and centromere

(メタゲノムおよびセントロメアにおける反復配列の
完全な再構成に向けて)

Yoshihiko Suzuki
鈴木慶彦

(Supervisor: Professor Shinichi Morishita)
(指導教員:　森下真一　教授)

A Dissertation

Submitted to
Department of Computational Biology and Medical Sciences
Graduate School of Frontier Sciences
The University of Tokyo

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Science

On December 11, 2019

Abstract

Sequence assembly, i.e., reconstruction of the original DNA sequences from their subsequences observed, has been one of the most principal challenges in bioinformatics since the dawn of its history. Repetitive sequences (*repeats*) are a major intrinsic source of the difficulty in assembling sequence fragments because they cause a serious ambiguity. Repetitiveness in a genome sequence takes several forms. The most typical repeats are interspersed repeats mainly caused by mobile genetic elements. Tandem repeats are abundant especially in eukaryotic genomes and known to be associated with many diseases and also as major components of centromeres in a variety of organisms. In a broad sense, ploidy of a genome is regarded as chromosome-scale repeats, and similar strains in a metagenome would be deemed as real-valued ploidy. The continuous evolution of sequencing technologies and algorithms for overcoming repeats has made it possible to achieve near-complete assembly of large genomes. Still now, however, there remain genomes and genomic regions difficult to perfectly reconstruct.

Various instances of the above-mentioned repeat types exist in real genomes. In this thesis, we especially focus on the following two relevant, unresolved research topics in genome science:

1) complete extrachromosomal mobile genetic elements (eMGEs), specifically plasmids and bacteriophages, in 12 human gut metagenomes; and
2) tandem repeats of ~360 bp units of the 1.688 gm/cm^3 satellite family in centromeres of *Drosophila melanogaster* F1 (A4×ISO strains) females.

In both cases, we assume no prior (incomplete) sequence databases and thus assemble the reads *de novo*. The types of the reads assembled are continuous long reads of the PacBio RS II sequencer with an error rate ~15% for metagenomes, and public circular consensus reads of the PacBio Sequel II sequencer with an error rate up to ~1% for centromeres.

A common strategy in genome assembly is the *overlap-layout-consensus* paradigm that first detects true overlaps between reads from all possible read pairs. We remark why the traditional pairwise method of overlapping reads

cannot resolve (approximate) repeats to rebuild the original sequences properly.

One can simply resolve approximate/exact repeats by spanning each repeat by reads. Finding a true overlap can be achieved with a high probability if the overlap is not contained in maximal approximate repeats of the genome sequence. One approach for locating unsettled repeats is to utilize some additional guidance on true read pairs such as optical mapping and Hi-C contact information. On the other hand, read overlap with a model on the background genome sequence is promising in finding more true overlaps by integrating out the sequencing errors on the conserved regions among repeats in the genome model. This provides a rationale for variant-detection-based methods employed in some recent works. There is, however, an inevitable limit due to sequencing errors where some of the pairwise read overlaps could fail even if we employ an appropriate genome model. Therefore, simultaneous estimation of the genome model and read overlaps is required ultimately.

Chapter 2: Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut

For the reconstruction of complete eMGEs from metagenomes, we took the approach based on external information. We first generated conservative *unitigs* and then mapped the labels of unitig binning onto the string graph to untangle edges on chimeric nodes by concatenating edges consistently labeled with the same bin ID.

De novo assembly of continuous long reads from 12 human faecal samples (~11 Gb total bases and ~8 kb mean subread length per sample) generated 82 eMGE contigs (2.5-666.7 kbp), which were classified as 71 plasmids and 11 bacteriophages, including 58 novel plasmids and six bacteriophages, and complete genomes of five diverse crAssphages with terminal direct repeats. In a dataset of 413 gut metagenomes from five countries, many of the identified plasmids were highly abundant and prevalent. Host microbes of the assembled eMGEs were predicted using several signatures including DNA methylation motifs reported by a PacBio's official tool, and the result suggested that Bacteroidetes-associated plasmids predominated, regardless of microbial abundance.

Chapter 3

第3章は5年以内に出版される予定であるため、最長で2025年3月22日までインターネット公表できません。
(Chapter 3 cannot be made public on the Internet until up to March 22nd, 2025, because the relevant content is scheduled to be published within 5 years.)

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Prof. Shinichi Morishita. His continuous assistance, guidance and advice both on my research and research style helped me very much for about six years and will continue to help. I am also deeply grateful to Prof. Gene Myers who has encouraged me to tackle with centromere sequence assembly since my first visit to his lab in 2017. He always surprises me with his deep insights and sophisticated algorithms on sequence alignment and assembly. I greatly appreciate everyone who kindly spared time for my doctoral degree including the reviewers of this thesis: Profs. Martin Frith, Takehiko Ito, Masahiro Kasahara, Shinichi Morishita and Yasubumi Sakakibara (listed alphabetically; the same applies hereinafter).

I am thankful to all the collaborators of my researches, Yoshikazu Furuta, Masahira Hattori, Shinichi Morishita, Gene Myers, Suguru Nishijima, Kenshiro Oshima, Wataru Suda and Jun Yoshimura (honorific titles omitted; the same applies hereinafter). They immensely widened the researches from various perspectives which could not be achieved without them.

I would like to appreciate those who gave me valuable comments and fruitful discussions: Bansho Masutani, Yuta Suzuki and other members in Morishita lab and Fabio Cunial, Martin Pippel and German Tischler-Höhle in Myers lab. I am strongly influenced by their ideas and works.

I would like to thank GSDM and JSPS for financial support. GSDM provided me the opportunity of the first visit to Prof. Myers lab in Germany for half a year, and JSPS enabled me to continue to visit the lab about every half year.

The contents in Chapter 2 have been published in *Microbiome* as an article (Suzuki et al., 2019). The texts, figures and tables are taken from the publication. I developed computational methods, performed computational analyses, and wrote the manuscript with Suguru Nishijima as a joint-first author. See the bibliography for details.

The contents in Chapter 3 have been under review in *Genome Research* as an article. I developed all methods, performed all analyses, and wrote the manuscript as the first author coauthored with Shinichi Morishita and Gene Myers.

Contents

Chapter 1

General introduction

Genome sequences are a fundamental basis of both biology and life sciences. However, for now we cannot observe an entire genome sequence at once, but instead observe fragments (i.e., *reads*) of the sequence with errors. Genome assembly is an inverse problem of reconstructing the original genome sequence from a set of these noisy substrings. Resulting sequences are seldom complete, and these are called *contigs* when no information on the gaps between them is available. Contigs with gap information are called *scaffolds*.

Note that there is a special case of the genome assembly problem where we already have previously assembled contigs or scaffolds and use them as a guide for the next assembly. This is called *reference-based assembly* or *resequencing* and useful especially for clinical applications, but we do not particularly touch the topic here. In other words, we treat only *de novo* genome assembly in this chapter.

The history of genome assembly is that of sequencing technologies and algorithms except the evolution of computers. As seen everywhere, there are some intimate combinations between specific sequencing technologies (i.e., data) and algorithms in genome assembly as well. In this chapter, we first outline the transition of the formulations of the genome assembly problem along with that of sequencing technologies. During it, so-called two major approaches currently used in genome assembly are introduced: the *string graph* and the *de Bruijn graph*, while the essence of both methods is quite same. Then, we go into an important subproblem generally employed in the former formulation: the *pairwise sequence alignment* problem. After that, we proceed to describe the difficulties in assembly of real genomes, specifically repetitive sequences (*repeats*), and related works for solving these issues.

1.1 Transitions of genome assembly

1.1.1 From shotgun sequencing to whole genome shotgun sequencing

The basic concept of shotgun sequencing is simple. Given multiple copies of a genome to be sequenced, it first shears the copies at random and then determines the sequences of the sheared fragments somehow. Additional biological experiments could be performed to determine the order and orientation of the contigs.

The effectiveness of genome assembly via shotgun sequencing using DNA cloning was practically demonstrated by the complete assembly of ~50-kbp bacteriophage lambda with Sanger sequencing (Sanger et al., 1982). Then, shotgun sequencing and its randomness were strongly supported by a statistical theory on the sequencing coverage and the number and length of contigs (Lander and Waterman, 1988) and another theory on the relationship between coverage and consensus sequence accuracy (Churchill and Waterman, 1992).

The genome assembly problem was initially regarded as the *shortest common superstring* (SCS) problem (Peltola et al., 1984), and several approximate and heuristic algorithms for the NP-hard problem were proposed (Turner, 1989; Ukkonen, 1990; Kececioglu and Myers, 1995). However, Sanger reads of the order of hundred nucleotides were much shorter than typical repeats in real genomes, and it was revealed that the SCS formulation was inappropriate for repeats longer than reads. That is, such repeats are over-compressed in the contigs produced based on the SCS criterion (Kececioglu and Myers, 1995).

In lieu of it, the early concept of the string graph was first introduced by Myers (1995) along with the formulation of the genome assembly problem using the Kolmogorov-Smirnoff (KS) test statistic, which requires reads to be uniformly distributed in the contigs to avoid over-compression of repeats. This graph-based layout strategy can be regarded as a successor to the *overlap graph* (Kececioglu and Myers, 1995) originally developed under the SCS formulation.

The string graph approach inherited the hierarchical architecture consisting of three major steps, namely *overlap-layout-consensus* (OLC), proposed in the overlap graph approach. That is, one first determines overlaps between reads from all possible read pairs in a manner described later, then builds a graph from the overlaps and picks up paths in the graph somehow to produce contigs, and finally computes a consensus sequence of each contig using reads belonging to the contig. The polished statement of the string graph was subsequently published in 2005 (Myers, 2005). The idea of KS statistic was then replaced with *A-statistic* of a contig that evaluates the repetitiveness of the contig (Myers et al., 2000; Myers,

2005) and a network-flow-based graph traversal for finding a generalized Eulerian tour in a string graph (Myers, 2005).

The shotgun method had been adopted in many assembly projects (The *C. elegans* Sequencing Consortium, 1998; International Human Genome Sequencing Consortium, 2001), but the clone-by-clone sequencing employed in these projects was extremely laborious especially in the case of large and complex genomes. Therefore, a more economical approach, the *whole genome shotgun sequencing*, began to be common around 2000 (Myers et al., 2000; Venter et al., 2001) with the help of technology leap including *paired-end* sequencing (Roach et al., 1995; Weber and Myers, 1997).

The application range of whole genome shotgun sequencing is beyond single genomes. *Whole metagenome shotgun sequencing* was also conducted to investigate microbial communities in a culture-independent manner (Venter et al., 2004; Tyson et al., 2004). Note that, in metagenome assembly, we cannot estimate the repetitiveness of a contig in the same manner as single genome assembly. This is because the abundance of each microbe in a metagenome is highly uneven (Venter et al., 2004; Arumugam et al., 2011).

1.1.2 Whole genome shotgun assembly with massive short reads

In the same volume of a journal in which the initial string graph concept was published (Myers, 1995), another novel k-mer-based method for genome assembly was proposed (Idury and Waterman, 1995). This direction was further pursued by Pevzner et al. (2001) and formalized using the de Bruijn graph.

The de Bruijn graph approach aims at finding a generalized Eulerian tour as well as the string graph, but the way of constructing a graph is different. That is, a de Bruijn graph is made from every k-mer occurring in the given reads and exact $(k-1)$-mer suffix-prefix matches between the k-mers. These restrictions enable a very fast graph construction while avoiding the quadratic time complexity of the all-vs-all overlap step in the string graph approach. This computationally efficient property of the de Bruijn graph matched very well with the high-throughput short-read sequencing emerged around 2005, and it quickly became widely used including metagenomics.

Another important merit of using k-mers is that it is easy to combine with algorithms based on the multiplicity of k-mers such as error correction (Pevzner et al., 2001). Namiki et al. (2012) utilizes the difference in the k-mer frequencies to decompose the de Bruijn graph of metagenomic reads into subgraphs each of which represents single species. Note that, however, such k-mer-based algorithms can indeed be applied independently of the de Bruijn graph.

1.1.3 The advent of single-molecule long reads

Short reads were unfortunately not suitable for perfect *de novo* genome assembly as described later, and sequencers capable of both high throughput and much longer read length were desired. Such sequencers practically appeared in 2011 for the first time and have continued to develop to date. Note that formerly "long reads" indicated Sanger reads, but hereinafter we mean only single-molecule long reads by long reads. We do not mention here about *long-range* technologies such as Hi-C, 10X and Bionano, although these are useful for versatile problems.

Long-read sequencing technologies, especially of Pacific Biosciences (PacBio) (Eid et al., 2009) and Oxford Nanopore Technologies (ONT) (Jain et al., 2015), directly handle single DNA molecules without amplification. This contributes to advantages other than read length, such as elimination of GC bias (Chaisson et al., 2015) and detection of native DNA methylation (mainly with PacBio so far) (Flusberg et al., 2010; Suzuki et al., 2016).

However, both technologies have a high sequencing error rate of ∼15% compared to the conventional sequencers. In addition, several specific error patterns have been confirmed although such errors occur at random in reads except some of those in the ONT sequencer that might be resolved by subsequent improvements on basecalling methods (Weirather et al., 2017). This intrinsic nature has forced researchers to develop specialized algorithms capable of handling very noisy reads. Some of them are introduced in the next section.

Unlike short reads, the string graph has been generally considered to match better than the de Bruijn graph for long reads because it can explicitly leverage the long read length and can ignore heavy sequencing errors via approximate string matching employed in the overlap step (Chin et al., 2016; Koren et al., 2017). Most of the early long-read assemblers perform the all-vs-all read overlap twice where the first stage computes consensus of each read using an alignment pileup and the second stage indeed determines overlaps for a string graph. However, some recent assemblers, such as Flye (Kolmogorov et al., 2019) and MARVEL (Nowoshilow et al., 2018), adopt strategies of skipping the first stage.

Flye's approach is based on repeat classification from alignments between raw reads. It builds a *repeat graph* in which each simple path corresponds to a repeat segment or unique segment in the genome, and then resolves the repeat segments as much as possible using spanning reads and small sequence variations within repeats. MARVEL is designed to be able to handle huge genomes, e.g., the axolotl genome of 32 Gbp, and has several features. MARVEL avoids the read correction phase by initial *scrubbing* of reads, i.e., correction of low-quality regions, unremoved adapters, ligation chimeras, etc., via investigation of the alignment pileup for each read. This strategy was originally proposed in

DAMASKER (https://github.com/thegenemyers/DAMASKER) in 2016.

More recently, accurate long reads produced by the *circular consensus sequencing* (CCS) technology of PacBio have been demonstrated practically useful (Wenger et al., 2019). This method scans a single DNA molecule repeatedly in a single process and outputs the consensus sequence of these multiple raw reads as an accurate read. CCS can produce reads with an error rate less than 1% while accomplishing average read length of ~20 kb. The technology itself has existed since 2010 (Travers et al., 2010), but has become a realistic choice by virtue of recent great advancement on the read length. Another exclusive and important feature of CCS is accurate detection of native DNA methylation at the single molecule level (Beaulaurier et al., 2015).

1.2 Pairwise sequence alignment for genome assembly

The first and critical step of the string graph approach (and other OLC-based approaches) is to list up pairs of reads overlapping with a high sequence similarity. Naively this can be carried out by just performing pairwise alignments (below) for every pair of reads, but it is too inefficient and practically unacceptable because most of the pairs are false and thus have a very low similarity. In this section, we briefly introduce algorithms and heuristics for pairwise sequence alignments and applications to the overlap step of genome assembly.

1.2.1 Pairwise sequence alignment via total score maximization

There are two well-known classical algorithms for computing an optimal alignment between two sequences: the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970; Sankoff, 1972; Sellers, 1974) for the global alignment and the Smith-Waterman algorithm (Smith and Waterman, 1981; Gotoh, 1982) for the local alignment. An alignment is called an *overlap alignment* (or *dovetail alignment*) if neither sequence is clipped. Overlap alignments are used for the construction of a string graph in genome assembly.

The optimal alignment obtained by these algorithms depends on the scoring function and the score matrix among the alphabets appearing in the sequences. Other than the originally proposed linear cost function, several variations on the scoring function have been proposed such as the affine gap cost function (Gotoh, 1982) and the convex cost function (Waterman, 1984; Miller and Myers, 1988; Sedlazeck et al., 2018). Custom score matrices are used mainly for protein sequence comparison (Dayhoff et al., 1978; Henikoff and Henikoff, 1992).

The process of the pairwise alignment with different scoring functions and score matrices can be interpreted in a unified manner through the *pair hidden Markov model* (PHMM) (Durbin et al., 1998). That is, the probability of an alignment in the model corresponds to the score of the alignment in a deterministic approach (Frith, 2019). A PHMM model can incorporate the knowledge on the source of the sequences compared such as sequencing quality (Frith et al., 2010; Hamada et al., 2011). This type of probabilistic framework has been employed for mapping sequences to a genome or quantitating similarity among distinct sequences.

1.2.2 Pairwise sequence alignment via edit distance minimization

Deterministic pairwise alignments can be performed through minimization of the *edit distance* (Levenshtein, 1966) between two sequences. This formulation was first introduced by Sellers (Sellers, 1974), and "diff" algorithms by Myers et al. (Myers, 1986; Wu et al., 1990; Myers, 2014) are now widely used although the same algorithm was independently discovered by Ukkonen (1985) prior to them. These algorithms run faster when the two sequences compared are more similar. This property matches well with the seed-and-extend heuristics described below. There is also a stand-alone implementation of fast edit distance computation named edlib (Šošić and Šikić, 2017) that extended Myers' bit-vector parallelization algorithm (Myers, 1999).

1.2.3 The seed-and-extend heuristics

Given two sequences, a naive end-to-end comparison using dynamic programming or diff algorithms has a quadratic worst-case time complexity and is practically slow, considering the amount and length of especially long reads. Therefore, initial screening of the possible alignment space with fast exact string matching is generally used. That is, one first detects exact match positions of short strings called *seeds*, then approximate alignments are *extended* from both ends of the seeds. Therefore, this strategy is called *seed-and-extend* heuristics. Exact string matching can be performed fast with algorithms such as hash table, suffix array and/or Burrows-Wheeler transform. Note that, as the short-seed matching is likely to be false, some techniques such as chaining of seeds (Abouelhoda and Ohlebusch, 2005; Kasahara and Morishita, 2006), the banded alignment and/or X-drop heuristics (Altschul et al., 1997) are combined together.

The seed-and-extend paradigm emerged in FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990; Myers, 2013). However, seeds stemming from repeats confound the seed matching by producing a huge number of false matches. Prior repeat masking is one workaround, but LAST efficiently adjusts the sensitivity and specificity using variable-length seeds adaptive to the input sequences (Kiełbasa et al., 2011). LAST showed the best performance in mapping ONT reads (Jain et al., 2015).

Recent overlappers for noisy long reads still have been built on top of the seed-and-extend strategy although each of these long-read aligners has its own philosophy, of course. DALIGNER (Myers, 2014) searches for seed hits via merging of sorted lists of k-mers with their positions in reads. Since the lists are very large, a cache-coherent parallelization of radix sort is developed and used. Another important feature is *trace point*. That is, DALIGNER stores only matching positions between the two reads for every w bp position (w is an arbitrary integer) for each alignment. The entire alignment path can be quickly and efficiently restored from these trace points while saving memory very much. This trade-off between time and memory can adapt to different types of computational demands. MHAP (Berlin et al., 2015) and minimap2 (Li, 2018) employ the same technique called *locality sensitive hashing* for space-efficient screening of seeds and fast seed matching. Both of these methods hash each read into a set of k-mers picked up from the read by "good" (i.e., non-biased) hash function(s). Canu (Koren et al., 2017) improved MHAP's sensitivity, when frequent k-mers are discarded, by employing adaptive weighting of k-mers by analogy of TF-IDF, a technique used in text search. Technically, Minimap2 adopts SIMD acceleration of the banded alignment originally developed in minialign (https://github.com/ocxtal/minialign).

1.3 Genome assembly is a fight against repeats

1.3.1 Conditions for perfect genome assembly

In both of the string graph and de Bruijn graph, there is a generalized Eulerian tour representing the original genome sequence (in the case of single genome assembly) provided that the sequencing coverage is sufficient and we could remove all the sequencing errors. However, it does not mean we can always trace the path unambiguously. This is because of repeats longer than reads. Myers (2014) stated that the requirements for perfect assembly are as follows:

1) reads are sampled from the genome sequence at random;
2) sequencing errors occur at random; and
3) every maximal repeats are spanned by reads.

More specifically about the condition 3), ambiguity arises when there exist a pair of interspersed repeats or triplet repeats longer than reads (Bresler et al., 2013). However, the precise definition of repeats given sequencing errors has not been considered sufficiently in the context of genome assembly. In Chapter 3, we propose a rigorous definition of *approximate repeats* in terms of false positive/negative rates of overlap detection, and remark why the traditional method of overlapping reads based on pairwise sequence comparison cannot resolve approximate repeats to rebuild the original sequences properly, as claimed by Tischler-Höhle (2019) and others.

The conditions above look simple, but in reality, practical algorithms for (near-)perfect assembly are highly non-trivial, and many algorithms have been developed depending on the type of repeats so far.

1.3.2 Types of repeats in genomes

Repetitiveness in a genome sequence takes several forms. The most typical repeats are *interspersed repeats* mainly caused by mobile genetic elements. *Tandem repeats* are abundant especially in eukaryotic genomes and known to be associated with many diseases (e.g., a recent publication on noncoding CGG expansions by Ishiura et al. (2019)) and also as major components of centromeres in a variety of organisms. In a broad sense, *ploidy* of a genome is regarded as chromosome-scale repeats, and similar strains in a metagenome would be deemed as real-valued ploidy. The continuous evolution of sequencing technologies and algorithms for overcoming repeats has made it possible to achieve near-complete assembly of large genomes, but there still remain many genomes and genomic regions difficult to perfectly reconstruct.

Various instances of the above-mentioned repeat types exist in actual genomes. In this thesis, we especially focus on the following two relevant, unresolved research topics in genoem science:

1) complete extrachromosomal mobile genetic elements (eMGEs), specifically plasmids and bacteriophages, in human gut metagenomes (in Chapter 2; with PacBio raw reads and short reads); and
2) tandem repeats of ~360 bp units of the 1.688 gm/cm^3 satellite family in centromeres of *Drosophila melanogaster* F1 females (in Chapter 3; with PacBio CCS reads).

1.3.3 Approaches to repeat assembly

Currently, there are no integrated approaches to repeat assembly. One can simply resolve approximate/exact repeats by spanning each repeat by reads (Bresler et al., 2013). Long-read technologies have dramatically alleviated the genome assembly problem in this manner so far. Note that, however, we should bear in mind that different assemblers produce different contigs, and a careful investigation and combination of results by multiple assemblers is usually required for complete assembly especially with noisy long reads in practice (Yoshimura et al., 2019).

One approach for locating unsettled repeats is to use some additional guidance on true read pairs, such as long-range technologies, other than continuous reads (Bishara et al., 2018; Kajitani et al., 2019). In metagenomes, sequence compositions and coverage can also be species-specific signatures for chimeric node resolution (Namiki et al., 2012). This strategy is employed in Chapter 2.

For more complicated regions, small sequence variations are usually used for separation of nearly-identical repeats: e.g., diploid phasing (Patterson et al., 2015; Chin et al., 2016; Kajitani et al., 2019), multi-class repeat separation (Tischler-Höhle, 2019; Bongartz, 2019), segmental duplication (Vollger et al., 2019), histone complex (Bongartz and Schloissnig, 2018), centromere (Jain et al., 2018). These methods will be mentioned in a little more detail in Chapter 3. On the other hand, however, these approaches are not directly applicable to highly divergent sequences, and a more careful approach is required as employed in Platanus-allee (Kajitani et al., 2019).

In Chapter 3, we propose a framework employing a *generative genome model* for detection of overlaps between reads more accurately than conventional pairwise alignments. As the first practical application of the framework, we developed a method for complex satellites using a Bayesian mixture model. Using simulation datasets, we demonstrated this framework and method indeed perfectly assembled long tandem repeats although state-of-the-art assemblers did not.

Chapter 2

Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut

Abstract

Elucidating the ecological and biological identity of extrachromosomal mobile genetic elements (eMGEs), such as plasmids and bacteriophages, in the human gut remains challenging due to their high complexity and diversity. Here, we show efficient identification of eMGEs as complete circular or linear contigs from PacBio long-read metagenomic data. *De novo* assembly of PacBio long reads from 12 faecal samples generated 82 eMGE contigs (2.5~666.7-kb), which were classified as 71 plasmids and 11 bacteriophages, including 58 novel plasmids and six bacteriophages, and complete genomes of five diverse crAssphages with terminal direct repeats. In a dataset of 413 gut metagenomes from five countries, many of the identified plasmids were highly abundant and prevalent. The coverage of gut plasmids by our plasmid data is more than twice that in the public database. Plasmids outnumbered bacterial chromosomes three to one on average in this metagenomic dataset. Host prediction suggested that Bacteroidetes-associated plasmids predominated, regardless of microbial abundance. The analysis found several plasmid-enriched functions, such as inorganic ion transport, while antibiotic resistance genes were harboured mostly in low-abundance Proteobacteria-

associated plasmids. Overall, long-read metagenomics provided an efficient approach for unravelling the complete structure of human gut eMGEs, particularly plasmids.

List of Abbreviations

ARG antibiotic resistance gene

CC circular contig

CO co-occurrence

COG cluster of orthologous groups

eMGE extrachromosomal mobile genetic element

IPD inter-pulse duration

IQR inter-quartile range

LMAG long-read metagenome-assembled genome

m6A 6-methyladenine

m4C 4-methylcytosine

MM methylation motif

PCC Pearson's correlation coefficient

POG phage orthologous group

SCC Spearman's correlation coefficient

SMRT sequencing Single Molecule, Real-Time sequencing

TDR terminal direct repeat

VLP viral particle

2.1 Introduction

Culture-independent metagenomics has provided a powerful approach to comprehensively explore microbial species and genes, which underlie an understanding of the ecological and biological features of the human gut microbiome (Gill et al., 2006; Kurokawa et al., 2007; Qin et al., 2010; The Human Microbiome Project Consortium, 2012). The metagenomes of microbial communities mainly comprise bacterial chromosomes and the associated extrachromosomal mobile genetic elements (eMGEs), such as plasmids and bacteriophages (phages). These eMGEs play important roles in microbial evolution, adaptation of the community to environmental changes, and interaction with hosts by conferring a variety of accessory functions on the community (Koonin and Wolf, 2008; Reyes et al., 2012; Virgin, 2014; Brito et al., 2016). For the analysis of plasmid communities (plasmidome), several specific procedures have been developed, including enrichment of closed circular plasmids by selective DNase treatment and CsCl-gradient ultracentrifugation from samples containing large amounts of linearized chromosomal DNAs (Dib et al., 2015; Jørgensen et al., 2015). For the bacteriophage community (phageome or virome), a crucial step is the enrichment of viral particles (VLPs) from samples containing vast numbers of microbial cells. VLP preparation requires several laborious techniques, such as stepwise filtration with different pore sizes and centrifugation under adjusted gravity conditions (Reyes et al., 2010; Minot et al., 2011, 2012, 2013; Castro-Mejía et al., 2015; Manrique et al., 2016; Shkoporov et al., 2018). However, these practices have not been well evaluated with respect to the quality and quantity of output data regarding the whole community structure.

It is also challenging to perform metagenomic sequencing of eMGE-enriched samples with short-read sequencers (Illumina and Ion Torrent) that can produce reads of only <500 bp. For example, *de novo* assembly of short reads generates notably short linear contigs (Qin et al., 2010; The Human Microbiome Project Consortium, 2012), possibly due to existing similar sequences among eMGEs and between eMGEs and chromosomes in a community. Such insufficient assembly makes it difficult to reconstruct full eMGEs as circular contigs (CCs), a structural hallmark of eMGEs excepting rare linear plasmids from metagenomic data, though there have been informatics tools that further connect the contigs to generate large bins (Qin et al., 2010; Nielsen et al., 2014). Therefore, most metagenomic studies based on short reads have analysed the whole community structure with little emphasis on separating microbial chromosomes and eMGEs (Li et al., 2014; Nishijima et al., 2016).

In contrast, long-read sequencers (Pacific Biosciences and Oxford Nanopore Technology) can produce long reads of ~10 kb or more. *De novo* assembly of long reads facilitates the generation of longer contigs and bins than those of short-read assembly by distinguishing among similar sequences (Sharon et al., 2015; Kuleshov et al., 2016; Brown et al., 2017; Frank et al., 2016; Tsai et al., 2016; Bishara et al., 2018). In addition, PacBio long-read metagenomics can also provide links between detected plasmids and their hosts using DNA methylation information (Beaulaurier et al., 2018). However, to date, there have been no intensive long-read metagenomic studies of eMGEs, indicating that human gut eMGEs remain to be explored. Therefore, we performed long-read metagenomics of whole faecal DNA samples to efficiently recover eMGEs as complete CCs from the assembled contigs and evaluated the diversity in human gut plasmids in this study.

2.2 Methods

2.2.1 Subjects, samples, and faecal DNA preparations[1]

We recruited 12 Japanese volunteers, of whom six subjects were the same as those who donated faecal samples in a previous study (Nishijima et al., 2016) and six other subjects were members of a family: two parents, two children, and two grandparents. No subjects were treated with antibiotics during faecal sample collection.

Faecal samples were transferred under anaerobic conditions at 4°C to the laboratory within 36 hours after defecation, immediately frozen with liquid nitrogen, and stored at -80°C until use. We collected 13 faecal samples from the 12 individuals, including a second sample (biological replicate) from an individual (denoted by ES) 2 months after the collection of the first sample. High-molecular-weight DNA samples were prepared by the enzymatic lysis method (Kim et al., 2013; Ueno et al., 2011). Prior to DNA extraction, each faecal sample suspended in PBS buffer was filtered with a 100-μm-mesh nylon filter (Corning Inc., New York, NY, USA) to remove human and eukaryotic cells and other debris from the faecal sample. The debris on the filter was washed twice using a glass or plastic bar with PBS buffer. The bacteria-enriched pellet was obtained by centrifugation of the filtrate at 5000 rpm for 10 minutes at 4°C (Ueno et al., 2011).

2.2.2 Sequencing of faecal DNA samples[2]

For SMRTbell library preparation, faecal DNA was sheared using a g-TUBE device (Covaris Inc., Woburn, MA, USA) at 4300 rpm and purified using a 0.45× volume ratio of AMPure beads (Pacific Biosciences, Menlo Park, CA, USA). SMRTbell libraries for sequencing were prepared using the “20-kb Template Preparation using BluePippin™ Size Selection System (15-kb Size Cutoff)” protocol. Briefly, the steps included 1) DNA repair, 2) blunt ligation with hairpin adapters with the SMRTbell template Prep Kit 1.0 (Pacific Biosciences), 3) 7-kb size cutoff size selection using the BluePippin DNA Size Selection System by Sage Science, and 4) binding to polymerase P6 using DNA Sequencing Reagent 4.0 (Pacific

[1] Sample preparation described in this subsection was conducted not by the author but by collaborators in Hattori laboratory (Waseda University, etc.).

[2] DNA sequencing described in this subsection was conducted not by the author but by collaborators in Hattori laboratory (Waseda University, etc.) and Morishita laboratory (The University of Tokyo).

Biosciences). Note that the DNA size selection might exclude some portion of the eMGEs in the samples although, as a result, we obtained circular eMGEs including those shorter than the cutoff length. SMRTbell libraries were sequenced on SMRT Cells (Pacific Biosciences) using magnetic bead loading and P4-C2 or P6-C4 chemistry. Sequence data were collected according to the magnetic bead collection protocol, 10-kb insert size, stage start, and 360-min movies in PacBio RS Remote. Primary filtering was performed on the PacBio RS II Blade Center server. The sequences mapped to the human genome (hg19) were removed prior to submission of PacBio reads to the NCBI Sequence Read Archive (SRA) using DAMAPPER (`https://github.com/thegenemyers/DAMAPPER`), a modified version of DALIGNER (Myers, 2014).

For short-read sequencing of seven newly collected samples in this study with the MiSeq platform, DNA libraries were prepared using the SPARK DNA sample Prep Kit (Qiagen, Beverly, MA, USA). Quality control of the metagenomic reads was conducted as described previously (Nishijima et al., 2016). Briefly, low-quality bases and reads were filtered using the FASTX tool kit (`http://hannonlab.cshl.edu/fastx_toolkit/`). Host-derived reads were excluded by mapping the reads to the reference human genome (hg19) using Bowtie2 (v.2.2.1) software (Langmead and Salzberg, 2012). The ratio of reads mapped to the human genome was <0.1% in both the long- and short-read sequencing (Supplementary materials: Table S2.1). The very low ratio of human reads in our metagenomic data can be explained by the efficient removal of human cells from the faecal samples by filtration prior to DNA extraction (Ueno et al., 2011), as described above. Additional metagenomic short reads (Roche 454, Ion PGM, and Illumina MiSeq) publicly available from the five countries (Li et al., 2014; Nishijima et al., 2016) were downloaded from the NCBI SRA.

2.2.3 Assembly of PacBio reads and short reads

For assembly of the PacBio metagenomic reads, we used FALCON v0.2 software (`https://github.com/PacificBiosciences/FALCON`) (Chin et al., 2016). Because FALCON tended to extend contigs to merge DNA sequences from distinct microbial species to generate erroneous contigs, we used unitigs, basic blocks of contigs that are shorter but more reliable contiguous sequences than contigs.

To reconstruct circular contigs (CCs) after FALCON assembly, we used the binning results of MetaBAT (Kang et al., 2015) as external guiding information with a single criterion: if a node in the assembly graph had only one in-edge and one out-edge that belonged to the same MetaBAT bin ID, then we merged the two edges representing unitigs to generate circular contigs. Note that a distinct bin ID was assigned to each unbinned unitig to avoid self-loops in the graph. This is the first attempt to map external binning information onto an assembly graph to untangle chimeric nodes in the graph. This method achieved reliable elongation of contigs by using the binning information to produce a more conservative layout of contigs than the original FALCON assembly did. To reconstruct relatively small circular contigs representing eMGEs, we used the cutoff values 2000 bp and 2200 bp for overlaps between raw subreads and between error-corrected subreads (technically, "preads"), respectively. These parameters influence the minimum length of the CCs generated by the assembly. After polishing the contigs with long reads using Quiver from the SMRT Pipe (v.1.87) software, the standard pipeline provided by Pacific Biosciences, we further corrected errors in the contigs using Pilon (v.1.12) (Walker et al., 2014), a software for error correction by short reads. The read depth of the assembled contigs was determined by PacBio's standard software. *De novo* assembly of the metagenomic short reads (Roche 454, Ion PGM, and Illumina MiSeq) was performed by MEGAHIT (v1.1.1) (Li et al., 2015).

2.2.4 Alignment of PacBio and short-read contigs

PacBio and short-read contigs were aligned using NUCmer (v3.1) software. Alignments with length coverage <95% or sequence similarity <95% were removed, and then, the sequence similarity of the alignments was calculated.

2.2.5 Estimation of microbial composition from PacBio and MiSeq data

To obtain the microbial composition from the PacBio data, we first predicted protein-coding genes in the PacBio contigs using Prodigal software (Hyatt et al., 2010). The genes were aligned to the 6149 reference genomes (Nishijima et al., 2016) using BLASTN with a >95% identity and >90% length coverage to assign the taxa (Arumugam et al., 2011). The relative abundance of the genomes/taxa was calculated by counting the number of genes aligned, multiplying the number of genes by the read depth of the contig, and normalising by gene length. Estimation of microbial composition from the MiSeq data was conducted by mapping the reads to the reference genomes using Bowtie2 with a 95% identity threshold and normalising the number of mapped reads by genome size (Nishijima et al., 2016). The similarity between the microbial compositions obtained from PacBio and MiSeq data was assessed with Pearson's correlation coefficient (PCC).

2.2.6 Reconstruction and analysis of HQ chromosome bins

For reconstruction of chromosome bins from the PacBio contigs in the 12 JP samples, metagenomic short reads (10 M reads per sample) of 106 JP individuals (Nishijima et al., 2016) were mapped to PacBio contigs by Bowtie2. Based on read depth and tetranucleotide frequency, contigs were clustered to chromosome bins using MetaBAT (v.0.26.3) (Kang et al., 2015) with the `--minMapQual 4 --verysensitive` options. The completeness and contamination were calculated by the presence or absence of single-copy marker genes using CheckM (v.1.0.5) (Parks et al., 2015), and high-quality (HQ) chromosome bins with >90% completeness and <5% contamination were defined. We deposited the sequences of 101 HQ chromosome bins tagged with the "long-read metagenome-assembled genome (LMAG)" in a public database (Supplementary materials: Table S2.7).

Taxonomic assignment of the HQ chromosome bins was conducted as previously described (Antipov et al., 2016). Briefly, the protein-coding genes predicted by Prodigal were aligned to 40 single-copy marker genes using BLASTP with an *E*-value <0.00001. The marker genes identified in the HQ chromosome bins were then aligned to those of the reference genomes using glsearch (v.36.3.5e) (Pearson and Lipman, 1988). The HQ chromosome bins having length-weighted average identity ≥96.5% with the reference genomes were assigned the same taxa as the reference genomes.

The phylogenetic tree of 101 HQ chromosome bins and 181 reference genomes with ≥0.05% relative abundance in the 12 subjects was constructed based on the similarity of amino acid sequences of the 40 marker genes using the neighbour-joining method in MEGA (v.6.06) (Tamura et al., 2013) and visualised with iTOL (Letunic and Bork, 2016). The similarities of the marker genes were calculated by MAFFT (v.7.043b) (Katoh and Standley, 2013) with the `--localpair --maxiterate 1000` options.

2.2.7 Classification of CCs as plasmids and phages

In the classification assessment using phage orthologous groups (POGs) (Kristensen et al., 2011), we determined the threshold of identity and length coverage to perform the highest confidence (Supplementary materials: Figure S2.2) using reference phages (n = 1957) as positive data and reference plasmids (n = 6589) as negative data available from NCBI on June 2016. By aligning the genes to POGs with BLASTP, the threshold (>90% length coverage) for classification of CCs as phages was determined. For classification of CCs as phages, VirSorter (v1.0.3) (Roux et al., 2015a) was also employed with the virome database and default options in the CyVerse environment (Merchant et al., 2016). Categories 1, 2, 4, and 5 were considered to classify CCs as phages, while categories 3 and 6 were excluded because these categories included false positives (Paez-Espino et al., 2016). PlasFlow (v1.1) was used with the default options for classification of CCs as plasmids (Krawczyk et al., 2018).

Functional annotation of genes in the CCs was conducted using Prokka (Seemann, 2014) and the Clusters of Orthologous Groups (COG) database (BLASTP with the *E*-value <0.00001). The presence and absence of known plasmid-enriched COGs related to plasmid replication, toxin-antitoxin system, and type IV secretion system (COG1475, COG2026, COG2126, COG2336, COG2948, COG3077, COG3451, COG3505, COG3704, COG3736, COG3843, COG5527, and COG5655) were investigated for CCs.

A similarity search of CCs for the public plasmid/phage database and phage sequences in the IMG/VR (Paez-Espino et al., 2016) and VirSorter (Roux et al., 2015a) databases was conducted using NUCmer (Kurtz et al., 2004), in which CCs with sequence similarity ≥90% and length coverage ≥70% to the references were assigned to the corresponding plasmids and phages, respectively.

The whole sequence comparison of the 71 plasmid CCs and 114 known/reference plasmids relatively abundant in the human gut was performed using TBLASTX (Mizuno et al., 2013). The 114 known/reference plasmids used in this analysis had average mapped reads of >5 per kb in the IGCJ dataset. The obtained dendrogram was visualised using iTOL software (Letunic and Bork, 2016).

2.2.8 Analysis of crAssphage genomes

PacBio subreads and MiSeq reads were aligned to the five CCs assigned to crAssphage. To assess the alignments, they were visualised using IGV (Thorvaldsdóttir et al., 2012). The sequences of the terminal direct repeats (TDRs) of the five CCs were obtained by reassembling subreads starting/ending at either side of the TDRs. MiSeq reads were further aligned to the TDR sequences using Bowtie2 to manually determine the exact ends of TDRs. To convert the circular genome of the crAssphage (NC_024711) in GenBank (Dutilh et al., 2014) to a linear genome, the TDRs were determined by aligning the TDRs of the five crAssphage CCs to the circular genome with BLASTN. Protein-coding genes in the linear crAssphage genomes were predicted using MetaGeneMark (Zhu et al., 2010), and the conserved genes in the six crAssphage genomes were investigated using Roary software (Page et al., 2015) with the `-p 80` option. The structures of the six crAssphage genomes were visualised using the genoPlotR package (Guy et al., 2010) in R software and custom Perl scripts. GC skew was calculated for a 100-bp sliding window with a 50-bp step size.

2.2.9 Quantification of eMGEs including the 82 CCs in the IGCJ dataset

We obtained all metagenomic reads from a total of 413 healthy faecal samples of Japanese ($n = 106$) (Nishijima et al., 2016), Danish ($n = 84$) and Spanish ($n = 59$) (Qin et al., 2010; Li et al., 2014; Le Chatelier et al., 2013), American ($n = 90$) (The Human Microbiome Project Consortium, 2012), and Chinese ($n = 74$) (Qin et al., 2012) people from http://public.genomics.org.cn, HMP DACC (http://www.hmpdacc.org), and the NCBI SRA to construct the IGCJ dataset. This dataset did not include data from patients with inflammatory bowel disease and type 2 diabetes. The metagenomic reads in the IGCJ dataset were subjected to quality control under the same conditions as described previously (Nishijima et al., 2016).

The eMGE clusters composed of 563 plasmid and seven phage clusters were constructed as follows. The IGCJ metagenomic reads (10 M reads per sample) were first mapped to all the publicly available plasmids and the 71 plasmid CCs using Bowtie2 with a 95% identity threshold. The reads hit >3000 plasmids, from which plasmids with map coverages <60% were excluded (see the "Results" section). The 1162 plasmids with mapped coverages ≥60% were then clustered with a ≥90% identity, ≥70% alignment coverage, and ≥0.7 ratio of shorter to longer sequences using NUCmer to generate 563 plasmid clusters. The breakdown of the plasmid clusters was 509 clusters of known/reference plasmids alone, 47 clusters of the novel plasmid CCs alone, and seven clusters of both plasmid CCs and those similar to known plasmids (Supplementary materials: Table S2.11). Similarly, we obtained a cluster of crAssphages and six unique clusters from the 11 phage CCs. The mapping of 10 M metagenomic reads per sample to the eMGE clusters was conducted with a ≥95% identity. The number of reads mapped to the clusters was normalised to the length of the longest representative eMGE in the cluster.

2.2.10 Host prediction of eMGEs

For host assignment of plasmids by similarity search, plasmid CCs were aligned to 5353 draft genomes publicly available with NUCmer (Kurtz et al., 2004), and draft genomes having a ≥90% identity and ≥70% length coverage with the CCs were assigned as the hosts of the corresponding plasmids.

For co-occurrence (CO) analysis, we mapped metagenomic reads of the IGCJ dataset to reference genomes and eMGEs with a 95% identity threshold to obtain the abundance normalised by genome size. Spearman's correlation coefficients (SCCs) were then calculated for variance in the abundance of chromosomes and eMGEs across the samples, and the genomes having SCCs of ≥0.7 with the eMGEs were predicted to be putative hosts of the corresponding eMGEs.

For host prediction of phages by CRISPR spacer similarity, we used three datasets of host genomes: the public genome database, contigs with ≥500 bp generated from assembly of metagenomic reads in the IGCJ dataset using MEGAHIT (v1.1.1) (Li et al., 2015), and contigs generated from the assembly of PacBio subreads in the JP PacBio dataset using Pilercr (v1.06) (Edgar, 2007). CRISPR spacers (≥20 bp) in microbial genomes and contigs were detected using Pilercr with the default options. The detected CRISPR spacers were aligned to the phage genomes using BLASTN with the following options: `-e 1 -G 10 -E 2 -q 1 -W 7 -F F`; this served to identify microbial genomes and contigs containing CRISPR spacers with 0 or one mismatch and >95% alignment coverage between them. The microbial taxa of the genomes and contigs were determined by their alignment using NUCmer to the reference genomes with a ≥90% identity and ≥50% alignment coverage.

The PacBio SMRT system can detect modified bases, such as 6-methyladenine (m6A) and 4-methylcytosine (m4C), because inter-pulse duration (IPD) between neighbouring bases is likely to be longer when the first bases are modified (Mizuno et al., 2013), and the modification is detectable by monitoring the IPD ratios of modified bases to those of unmodified ones. According to the process described previously (Beaulaurier et al., 2018), we first determined the optimal parameters of "methylation fraction" (percentage of motif sequences methylated), "mean coverage" (average sequencing read-depth per strand on the motif sites), and "mean IPD ratio" to 0.6, 25, and 2.5 as the thresholds, respectively, from PacBio reads from a mock community composed of eight bacteria with and without plasmids (*Lactobacillus paralimentarius* JCM 10707, *Natronolimnobius baerhuensis* JCM 12253, *Bacillus cereus* ATCC 14579, *Variovorax* sp. JCM 16519, *Clostridiales* bacterium ACSP 3, *Staphylococcus aureus* HSAU10, *Bifidobacterium longum* IBLI, and *Escherichia coli* SE11). We then filtered for methylation motifs (MMs) in the HQ chromosome bins with the optimised methylation fraction and mean coverage. In this process, we excluded the motif G^{m6}ATC from host prediction because this motif was ubiquitous among bacteria. Using the filter-passed chromosomal MMs as baits, we calculated the mean IPD ratio values of the MMs in each eMGE and HQ chromosome bin and binarized the values according to the threshold (i.e., IPD ratios higher than the threshold were defined as 1 to indicate methylation, and the others were defined as 0 to indicate nonmethylation). Finally, we linked the eMGEs and the HQ chromosome bins, between which at least one MM was shared, and the binarized IPD ratio values were equivalent except missing values.

The results of host prediction of the plasmid CCs were summarised and visualised as a host-plasmid network using Cytoscape. In this analysis, taxonomically undefined bacterial species (e.g., *Bacteroides* sp.) were changed to taxonomically defined bacterial species of which the 16S rRNA gene sequence had ≥99.8% identity with that of the undefined species.

2.2.11 Comparison of functions between plasmids and chromosomes

For comparison of the frequency of COGs between plasmids and chromosomes, we used 315 relatively abundant plasmids (≥1 average mapped reads per 10 kb) and complete chromosomes of 249 microbial species with ≥0.1% average abundance in the IGCJ dataset. The genes were functionally annotated by BLASTP to the COG database with the E-value <0.00001 using Prodigal (Hyatt et al., 2010). Statistical significance was calculated using Fisher's exact test, and p-values were transformed to q-values (Storey and Tibshirani, 2003). Antibiotic resistance genes were identified by searching Resfams database (Gibson et al., 2015) using the `hmmscan` function of HMMER3 (Finn et al., 2011) with the gathering thresholds. The abundances of the ARG (antibiotic resistance gene)-positive and ARG-negative plasmids were compared using the Wilcoxon rank-sum test.

2.3 Results

2.3.1 Metagenomic sequencing of human faecal samples with the PacBio SMRT system

We sequenced 13 faecal DNA samples from 12 healthy Japanese adults, including one biological duplicate (ES1-2 and ES9-1). A total of ~11 Gb per sample with an average subread length of 8 kb was obtained from 10 individuals (excluding two subjects with poor subread lengths) with the PacBio RS II system. We also generated short reads from six of the 12 subjects with three short-read sequencers (Illumina, 454 and Ion PGM) and obtained them from a previous publication for the other six subjects (Nishijima et al., 2016). The sequencing statistics are summarised in Supplementary materials: Table S2.1.

We, therefore, conducted *de novo* assembly of the PacBio and short reads by using FALCON and MEGAHIT as assemblers, respectively (see the "Methods" section). We compared the two assembly outcomes from the data of three samples (apr34, apr38, and FAKO02) with similar sequence amounts in PacBio and short-read sequencing. The comparison revealed that PacBio reads boosted assembly statistics, with an N50 contig length reaching ~202 kb, while those of the short reads were ~4 kb (Fig. 2.1a). The results of the long-read assemblies showed that the N50 contig length ranged from 24.6 to 279.2 kb for all the samples (Supplementary materials: Table S2.2). We then evaluated the accuracy of the PacBio contigs based on the sequence similarity between PacBio and the corresponding short-read contigs of the same samples. The results revealed that PacBio contigs with 5, 10, 20, and ≥40 read depths were aligned with short-read contigs with 99.4, 99.7, 99.8, and ≥99.9% identities, respectively (Fig. 2.1b). Assuming the accuracy of the aligned short-read contigs to be sufficiently high, the accuracy of PacBio contigs with read depths >5 could be estimated to be >99.4%, accounting for ~99.8% of the total contig length (Supplementary materials: Table S2.3).

2.3.2 Microbial and gene composition in PacBio metagenomic data

We compared the microbial abundance estimated from the PacBio and MiSeq reads. Taxonomic assignment of PacBio data was performed by similarity search of genes predicted in PacBio contigs for the reference genomes, followed by counting the number of PacBio reads mapped to the genes to quantify their abundance (see the "Methods" section), while that of the MiSeq data was performed by direct mapping to the reference genomes as described previously (Nishijima et al., 2016). The estimated microbial abundances between the two data points in each subject were significantly similar at the genus level, with a median PCC of ~0.99, which was significantly higher than that among the 12 individuals (Fig. 2.1c, d).

The mean gene length in the PacBio contigs was 847 bp, longer than the 662 bp in the short-read contigs and closer to the 957 bp of mostly full-length genes in the reference genomes (Supplementary materials: Figure S2.1a). In addition, an average of 27.6 genes was identified per PacBio contig, which was ~10 times more than the 2.4 per short-read contig on average (Supplementary materials: Figure S2.1b).

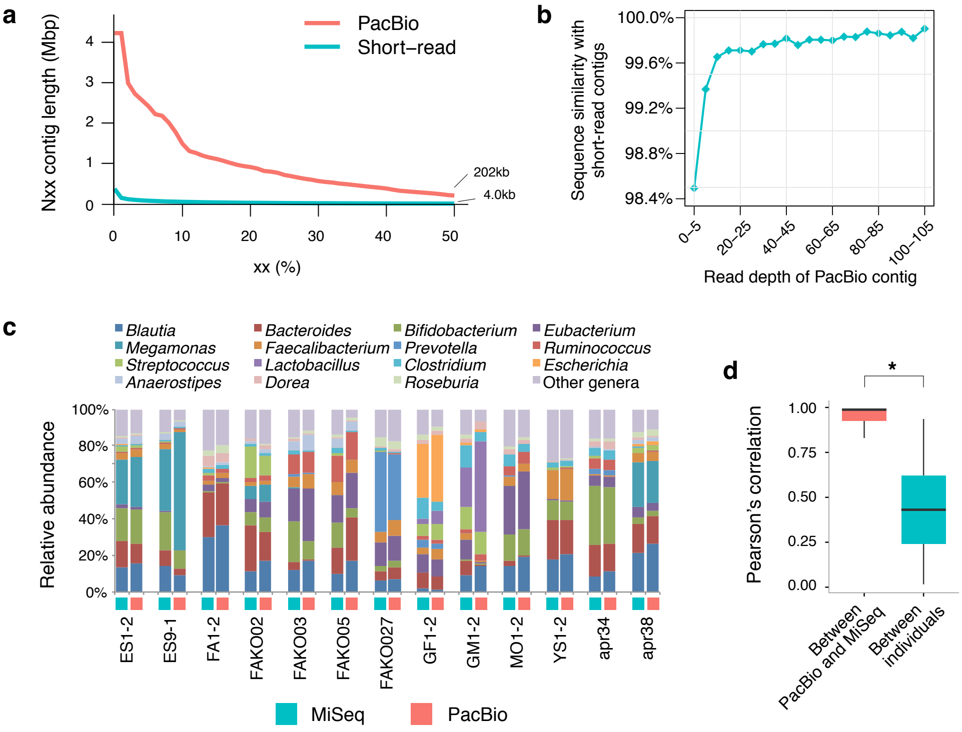
a
Nxx contig length (Mbp)
4
3
2
1
0
PacBio
Short−read
202kb
4.0kb
0
10
20
30
40
50
xx (%)
b
Sequence similarity with short-read contigs
100.0%
99.6%
99.2%
98.8%
98.4%
0–5
20–25
40–45
60–65
80–85
100–105
Read depth of PacBio contig
c
Blautia
Megamonas
Streptococcus
Anaerostipes
Bacteroides
Faecalibacterium
Lactobacillus
Dorea
Bifidobacterium
Prevotella
Clostridium
Roseburia
Eubacterium
Ruminococcus
Escherichia
Other genera
Relative abundance
100%
80%
60%
40%
20%
0%
ES1-2
ES9-1
FA1-2
FAKO02
FAKO03
FAKO05
FAKO027
GF1-2
GM1-2
MO1-2
YS1-2
apr34
apr38
MiSeq
PacBio
d
*
Pearson's correlation
1.00
0.75
0.50
0.25
0.00
Between PacBio and MiSeq
Between individuals

Figure 2.1: Statistics of metagenomic sequencing of 13 faecal samples with the PacBio SMRT system and short-read sequencers. **a** To show the length distribution of the contigs of long and short reads, we selected three samples (apr34, apr38, and FAKO02) that had similar sequence amounts in both PacBio long-read and short-read sequencing (see the "Results" section). The y-axis shows the Nxx contig length, an indicator of measuring the quality of genome assembly such that xx% of all bases in the assembled contigs of the three selected samples are found in contigs of the Nxx contig length or more, while the x-axis shows the value of xx, which measures coverage of bases by contigs. **b** Sequence similarity between PacBio and short-read contigs. The y-axis shows the sequence similarity of the PacBio contigs with the reciprocally best-matched short-read contigs, and the plots show the average value for every five units of read depth of the PacBio contigs on the x-axis. PacBio and short-read contigs of the 12 samples were aligned using NUCmer with a >95% identity and a >95% length coverage. **c** Genus-level microbial compositions estimated from the PacBio and MiSeq data of the 13 samples. Taxonomic assignment and quantification of microbial abundance from the PacBio and MiSeq data were described in the "Methods" section. **d** PCCs between the microbial compositions estimated from PacBio and MiSeq data. PCCs (left) between the same samples, excluding the biological replicates (ES1-2 and ES9-1), and PCCs (right) between different samples are shown. The boxes represent the inter-quartile range (IQR), and the lines inside represent the median. The whiskers show the lowest and highest values within 1.5 times the IQR. Asterisks represent $p < 0.01$ (Wilcoxon rank-sum test).

2.3.3 Circular contig generation from PacBio read assembly

In the assembly, we set the minimum overlap length between two subreads to 2200 bp (see the "Methods" section), though CCs smaller than the cutoff (2.2 kb) cannot be identified by this method. The assembly generated a total of 82 CCs ranging from 2.8- to 666.7-kb in length (Supplementary materials: Table S2.4). To test whether these CCs were eMGEs, we classified them as plasmids and phages using several classification assessments, such as searching POGs (Supplementary materials: Figure S2.2) (Kristensen et al., 2011), VirSorter (Roux et al., 2015a), and PlasFlow (Krawczyk et al., 2018), checking the presence or absence of known plasmid-enriched genes, such as mobilisation- and conjugation-related genes, and a similarity search of the public database. Because the POG and VirSorter assessments classified 11 CCs (30.2 to 98.9 kb in size) as phages with high consistency, we classified the remaining 71 CCs as plasmids (2.8 to 666.7 kb). A similarity search of the public plasmid/phage database revealed that 17 of the 71 plasmid CCs were highly similar to 10 known plasmids, and five of the 11 phage CCs were highly similar to a genome of a crAssphage, NC_024711.1 (Dutilh et al., 2014).

To further confirm the accuracy of the classifications, we blasted the CCs against the virome databases VirSorter and IMG/VR (Roux et al., 2015b; Paez-Espino et al., 2016). The five CCs assigned to crAssphage and a putative novel phage CC (FAKO05_000032F) hit several sequences in the virome databases, consistent with the present classification. However, five plasmid-classified CCs (FA1-2_2760, FAKO05_2268, FAKO05_2271, FAKO27_6410, and FA1-2_000589F) matched sequences in the virome databases (Supplementary materials: Table S2.4), showing disagreement with the present classification (see the "Discussion and conclusions" section).

We clustered the 71 plasmid CCs with 114 known plasmids relatively abundant in the human gut based on overall sequence similarity (Fig. 2.2a, see the "Methods" section). The results revealed that many of the 71 CCs had high sequence diversities for the known plasmids. Based on the host taxa of the known plasmids, most of the 71 CCs aggregated in Firmicutes and Bacteroidetes plasmids, and many of the novel CCs aggregated in Firmicutes plasmids, while only four novel CCs aggregated in Proteobacteria plasmids.

We also identified two highly similar, in terms of sequences, but distinct plasmid CCs in the assemblies of long reads from three subjects (apr34, FAKO03, and FAKO05). The two similar CCs in each subject had a sequence alignment of length >1 kb with >99% identity between them, but in the short-read assembly, either the corresponding sequences were fragmented into multiple contigs or only one of the two CCs was generated (Supplementary materials: Figure S2.3). These results demonstrated that similar plasmids hard to distinguish in short-read assembly can be precisely reconstructed as independent contigs in long-read assembly. Overall, we identified 82 CCs and classified them as 71 plasmids and 11 phages, of which 58 plasmid and six phage CCs are likely to be novel eMGEs (Supplementary materials: Figure S2.4).

We further performed the functional annotation of genes in the 71 plasmid CCs using the COG database. The data revealed that ~47% of the genes identified were novel, and genes assigned to COG category X, "Microbiome", were most enriched in the functionally annotated genes, as expected (Supplementary materials: Table S2.5).

2.3.4 Structure of contigs assigned to the crAssphage genome

Mapping of PacBio and short reads to the five crAssphage CCs suggested that these CCs had a linear genome with TDRs of length ~2 kb. This was supported by several lines of evidence, e.g., approximately twofold higher coverage of both PacBio and short reads mapped to the TDR region than other regions in the circular genome, absence of PacBio reads spanning the TDRs, and higher frequency of both PacBio and short reads starting from both ends of the TDRs than reads from other positions (Supplementary materials: Figure S2.5). Both TDRs in each genome were almost identical, while the sequence similarity and length slightly varied among TDRs in the five crAssphages (Supplementary materials: Table S2.6 and Figure S2.6). The linear genomes of six crAssphages, including NC_024711.1, encoded 89 to 91 putative genes, of which 61 were highly conserved with ≥80% amino acid identity among them; the number of genes unique to each genome ranged from 0 to 16 with an average of 6.3 per genome, and other conserved genes numbered between two and five (Fig. 2.2b). Additionally, the genomes exhibited a clear transition in GC skew of the coding strand at approximately 30 kb away from the right TDR (Supplementary materials: Figure S2.7). Similarly, two phage CCs (FAKO05_000032F and FAKO27_000271F) were found to have linear genomes by mapping the reads to the CCs (Supplementary materials: Figure S2.8). Our data indicated that linear phage genomes with TDRs were erroneously assembled as CCs. The TDRs are the source of this mis-assembly, which could be corrected by mapping the reads to CCs as described previously (Chung et al., 2017).

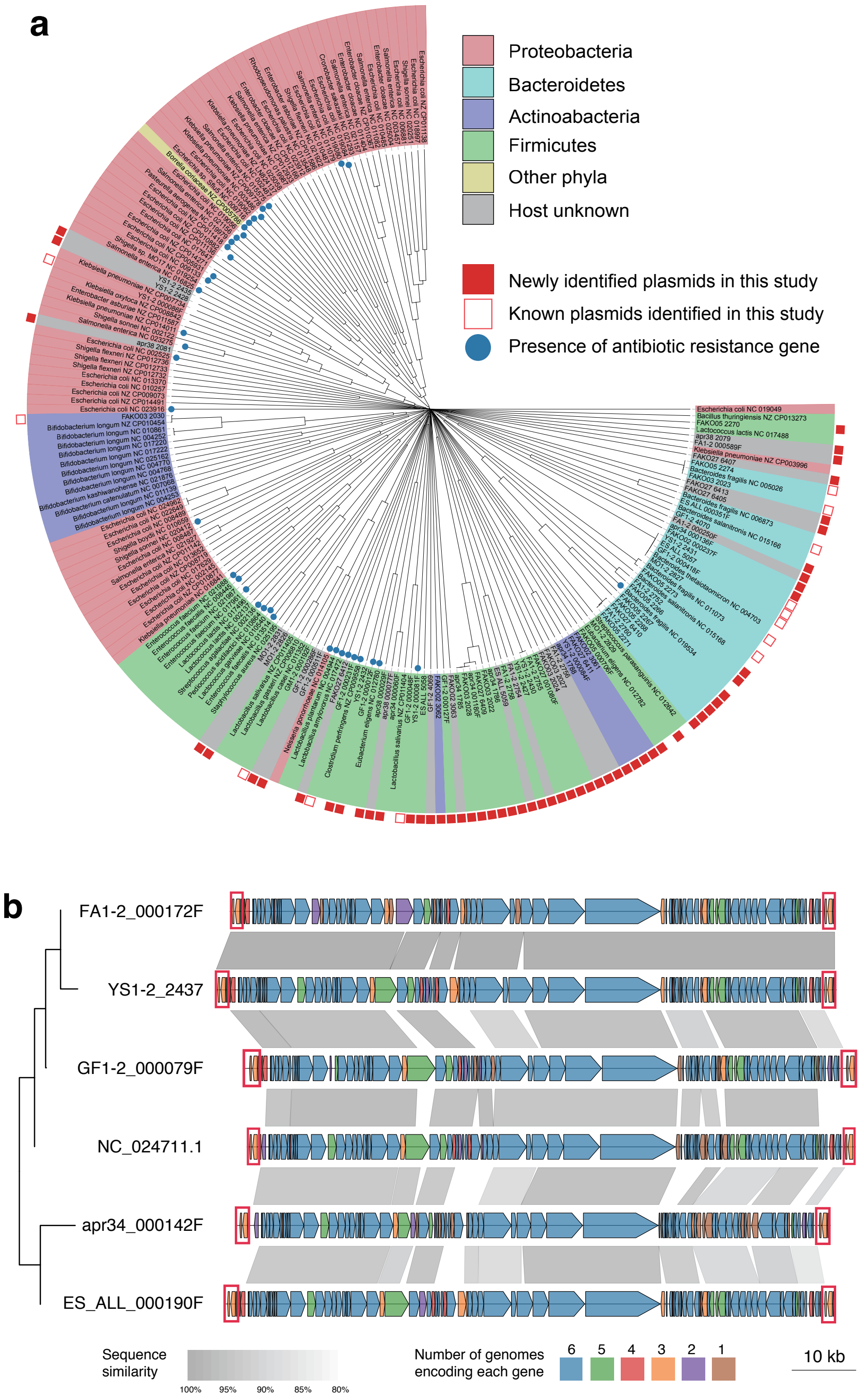
a
Proteobacteria
Bacteroidetes
Actinoabacteria
Firmicutes
Other phyla
Host unknown
Newly identified plasmids in this study
Known plasmids identified in this study
Presence of antibiotic resistance gene
b
FA1-2_000172F
YS1-2_2437
GF1-2_000079F
NC_024711.1
apr34_000142F
ES_ALL_000190F
Sequence similarity
100%
95%
90%
85%
80%
Number of genomes encoding each gene
6
5
4
3
2
1
10 kb

Figure 2.2: Whole-sequence comparison of 71 plasmid CCs and structure of six crAssphage linear genomes. **a** Dendrogram of 71 plasmid CCs and 114 known plasmids that were relatively abundant in the human gut (see the "Methods" section). The phyla are shown in different colours (green for Firmicutes, purple for Actinobacteria, red for Proteobacteria, blue for Bacteroidetes, yellow for other phyla and grey for unknown hosts). Red squares in the outer circle indicate the plasmid CCs newly identified in this study. Blue circles on the edges show the presence of antibiotic resistance genes. **b** Putative genes shown by pentagons in the linearized genomes of five crAssphages identified in this study and NC_024711.1 (Roux et al., 2015b). Each grey shade connecting two genomic regions indicates the average sequence similarity of the region. The left dendrogram shows a clustering of the six genomes based on overall similarity. To show the degree of conservation of each putative gene in the six genomes, six different colours are used. Brown genes are unique to only one genome, while blue genes are shared in common by all genomes. The red boxes at the ends indicate TDRs in the linear genomes.

2.3.5 Reconstruction of microbial chromosomes from PacBio contigs

The assembly of PacBio reads also yielded seven large CCs from 2 to 3 Mb in length, which were considered to be bacterial chromosomes. We additionally reconstructed 94 HQ chromosome bins (completeness >90%, contamination <5%) with putative genome sizes ranging from 1.88 to 6.83 Mb, in which multiple rRNA genes were consistently allocated (Supplementary materials: Table S2.7). Of these chromosome bins, 17 might be phylogenetically novel, because their identities with known genomes were lower than the threshold (96.5%) (Mende et al., 2013). Phylogenetic tree analysis indicated that 69 bins, including the 17 novel bins, were taxonomically classified as Firmicutes, 18 as Bacteroidetes, 13 as Actinobacteria, and one as Proteobacteria (Supplementary materials: Figure S2.9).

2.3.6 Host prediction of eMGEs

Host prediction of the 82 eMGEs was performed by several methods: sequence similarity search for publicly available draft genomes (Antipov et al., 2016), co-occurrence profile based on abundance (CO) (Dutilh et al., 2014), methylation motif (MM) similarity (Beaulaurier et al., 2018), and CRISPR spacer similarity to only the phage's host (Stern et al., 2012; Edwards et al., 2015).

A similarity search of the 71 plasmid CCs for the draft genomes showed that 36 CCs hit the draft genomes of various strains, which were taxonomically well-matched with those assigned by the similarity search for known plasmids (Supplementary materials: Table S2.8 and S2.9). In the host prediction by CO analysis, we used the IGCJ dataset composed of 413 faecal metagenomic data from Japan (JP), the US (US), Spain (ES), Denmark (DK), and China (CN) (see the "Methods" section) (Li et al., 2014; Nishijima et al., 2016). We identified nine CCs that had SCCs (Dutilh et al., 2014) of >0.7 for variance in abundance with several genomes/hosts across the samples (Supplementary materials: Table S2.9). The MM similarity search using the present JP PacBio dataset found 19 plasmid CCs that shared 26 different MMs with 14 HQ chromosome bins (Supplementary materials: Figure S2.10 and Table S2.9).

As shown in Fig. 2.2a, many of the plasmids, including the host-predicted plasmid CCs, tended to be grouped by host taxa, except for the five Actinobacteria-predicted novel CCs that segregated from the known Actinobacteria plasmids.

We further constructed a host-plasmid network from the host-predicted plasmid CCs and found many shared plasmids between various *Bacteroides* species and several *Parabacteroides* and *Prevotella* species, forming a large network distinct from others in the human gut microbiomes of the 12 subjects (Supplementary materials: Figure S2.11).

In the host prediction of phages, because no host candidate was identified in the CO analysis and the similarity search, we used three different datasets (JP PacBio, IGCJ, and the public genome database) for CRISPR spacer similarity search and the JP PacBio dataset for the MM similarity search. Four phage contigs (FAKO27_000271F, YS1-2_2434, FAKO27_000238F, and apr34_1784) had nearly perfect matches with CRISPR spacers in several genomes of the three datasets (Supplementary materials: Table S2.10) and concurrently shared 13 MMs with four genomes in the JP PacBio dataset (Supplementary materials: Figure S2.10). The hosts of the four phages as predicted by the two methods were consistent taxonomically. In the host prediction of seven other phage contigs by CRISPR spacer similarity alone, six including the five crAssphages had similarity to CRISPR spacers in the genomes of *Bacteroides* and *Porphyromonas*, both of which belong to the order *Bacteroidales*, in at least two datasets. The host for one phage (apr34_1792) was predicted to be *Bifidobacterium* in only the IGCJ dataset (Supplementary materials: Table S2.10). Overall, hosts for 50 plasmid and 11 phage CCs were predicted, while no host was predicted for 21 plasmid CCs by the methods used. In this host prediction, we cannot exclude the possibility that hosts of eMGEs can also be extended to phylogenetically different taxa close to the predicted tax.

2.3.7 Quantification of gut eMGEs using 413 metagenomic datasets from five countries

For quantification of gut eMGEs in the IGCJ dataset, we constructed and used eMGE clusters composed of 563 plasmid and seven phage clusters to which the IGCJ metagenomic reads were mapped. For construction of the eMGE clusters, we first mapped all the plasmids publicly available by IGCJ metagenomic reads with a ≥95% identity and excluded the plasmids with mapped coverage <60% because many of them included plasmids unevenly mapped by non-specific reads containing conserved genes such as transposases and very low-abundance plasmids that were considered to be negligible for quantification. Clustering of the plasmids with mapping coverage ≥60%, 11 phage CCs and all publicly available crAssphages generated the eMGE clusters, each of which was composed of highly similar eMGEs with a ≥90% sequence identity and ≥70% alignment coverage. Mapping of 10 million (M) short reads per sample to these eMGE clusters revealed that ~1.1% of the total reads on average were mapped to the plasmid clusters and ~0.38% to the crAssphage cluster (Fig. 2.3a and Supplementary materials: Table S2.11). Our novel plasmid CCs accounted for ~60% of the total reads mapped to the plasmid clusters, indicating that many of them were highly abundant in the

IGCJ dataset (Fig. 2.3b). The inter-country variability in the average abundance of crAssphages (0.03 to 1.4%) was remarkable compared with that of plasmids (0.56 to 1.54%) (Fig. 2.3a and Supplementary materials: Figure S2.12a and Table S2.11). The increased abundance of crAssphages in the US dataset was largely due to the existence of several subjects having extremely high-abundance crAssphages (up to ~21%) but not due to extensive prevalence (Supplementary materials: Figure S2.12b and c). Indeed, the proportion of crAssphage-positive subjects in the US dataset was ~53%, slightly lower than the average (~60%) of the five countries (Supplementary materials: Figure S2.12c).

In the top 20 highly abundant eMGE clusters, 12 including the top four plasmid clusters were associated with Bacteroidetes as putative hosts (Fig. 2.3c). Likewise, analysis of the host taxon distribution of plasmids revealed that Bacteroidetes-associated plasmids had higher abundance than plasmids associated with other phyla (Fig. 2.4a). This Bacteroidetes dominance was observed in all the countries, varying from a minimum of 61% in the JP dataset, with 17% Bacteroidetes, to a maximum of 93% in the US dataset, with 66% of the total microbial abundance representing Bacteroidetes (Fig. 2.4b). The top 20 eMGE clusters included two phage clusters (crAssphage [Cluster_F1] and Bacteroides phage [Cluster_F2]). Notably, the latter (FAKO05_000032F) had higher average mapped reads than the crAssphages in the DK dataset and slightly higher average prevalence (~71%) than the crAssphages (~60%) in the IGCJ dataset (Supplementary materials: Table S2.11).

We next estimated the ratio of gut plasmids and crAssphages to microbial cells for each of the five countries. The estimation was based on the number of reads mapped to the plasmid and crAssphage clusters and the average sizes of microbial chromosomes, plasmids, and crAssphages. The results revealed that the average ratio of eMGEs to microbial chromosomes ranged from 1.2 to 4.3 for plasmids (3.0 on average) and from 0.01 to 0.7 for crAssphages (0.18 on average) (Supplementary materials: Table S2.12). These data showed that gut plasmids outnumbered microbial cells on average, but crAssphages did not outnumber the microbial cells in the IGCJ dataset. Only in the US dataset were crAssphages close in number to microbial cells, with an average ratio of 0.69. There was no significant correlation between the abundance of crAssphages and subjects' age, BMI, and sex (Supplementary materials: Figure S2.13).

2.3.8 Functional profiles of gut plasmids in 413 metagenomic datasets

Functional annotation of 315 plasmids and 249 chromosomes relatively abundant in the IGCJ dataset revealed that 360 COGs had significant differences (*q*-values

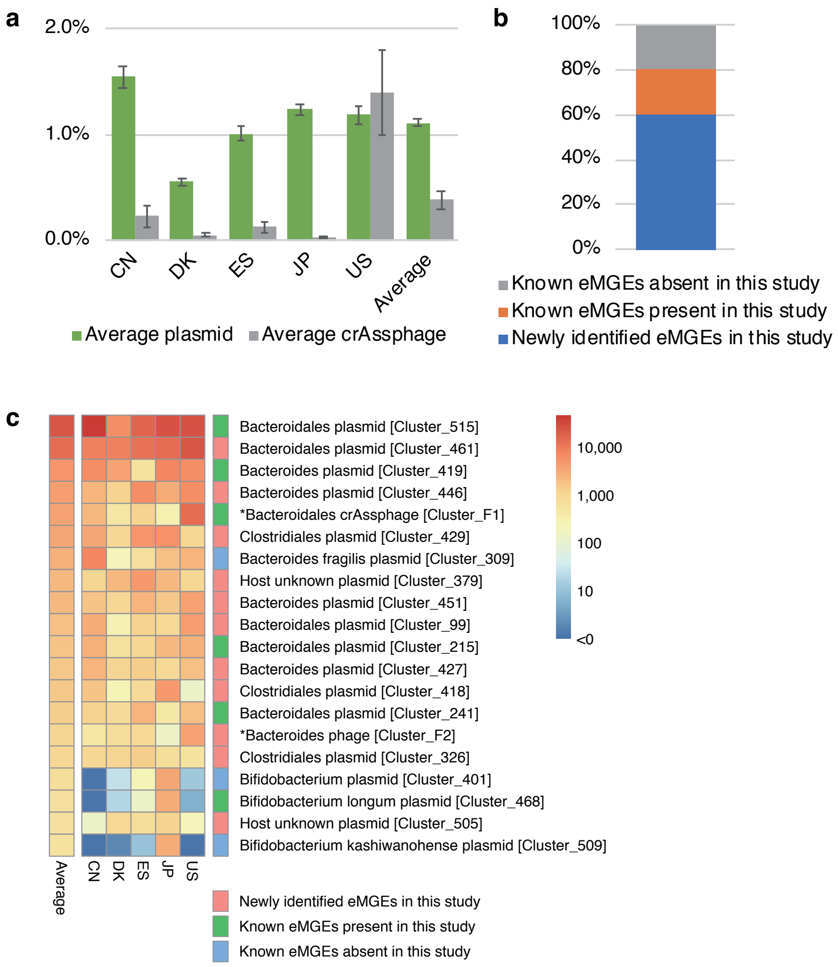

Figure 2.3: Quantitative analysis of eMGEs in the IGCJ dataset. **a** Average ratios of metagenomic reads mapped to non-redundant eMGE clusters. Error bars represent standard mean errors. **b** Average ratios of reads mapped to three classes of eMGEs. Newly identified eMGEs, known eMGEs present, and known eMGEs absent in this study are shown by blue, orange, and grey, respectively. **c** Heatmap of the abundance of eMGE clusters in the IGCJ dataset. The abundance is the number of reads mapped to eMGEs normalised by length. Colour shades show the degree of abundance of the eMGEs; red indicates relatively high abundance, while blue indicates relatively low abundance. Three classes of eMGEs are also shown by three colours, blue, green, and red, respectively (also see Supplementary materials: Table S2.10).

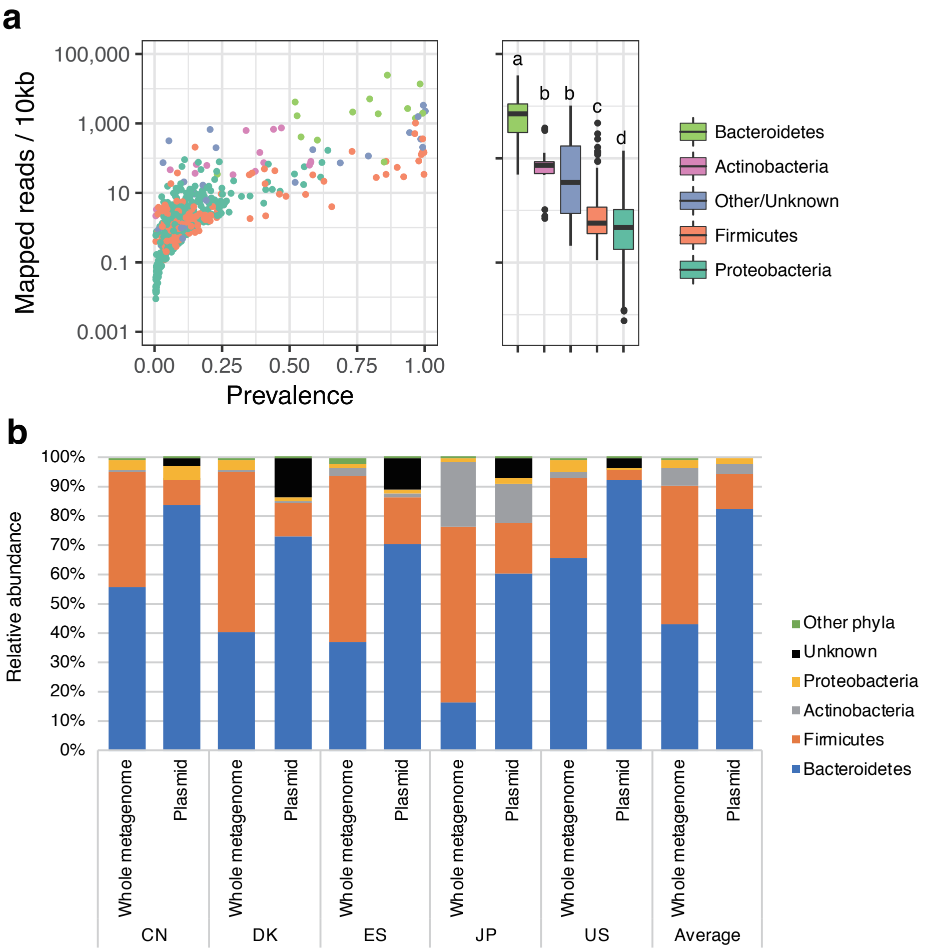

Figure 2.4: Taxonomic distribution of plasmid-associated hosts in the IGCJ dataset. **a** Abundance and prevalence of plasmid-associated hosts at the phylum level in the IGCJ dataset. The left dot plot shows the abundance (y-axis) and the prevalence (x-axis) of each plasmid. Putative hosts are assigned to plasmids and are grouped into four major taxa, and unknown and other taxa are coloured differently. The right box plot shows the abundance distributions of the phylum of plasmid hosts depicted by the IQR and median. The whiskers show the lowest and highest values within 1.5 times the IQR. The letters (a, b, c, and d) above the boxes indicate statistically significant ($p < 0.01$) differences between phyla with different letters. **b** Phylum-level compositions of host taxa from the whole metagenome data and plasmids in the five countries. The average abundance of the phyla, depicted by different colours, in the five countries is shown.

< 0.05) in abundance between them, and 233 COGs were significantly enriched in plasmids (Supplementary materials: Table S2.13, see the "Methods" section). In particular, eight were detected only in the plasmids; two were related to inorganic ion transport (COG4264 and COG2370), one was a type IV secretory pathway VirB6 component (COG3704), and the remaining five were uncharacterized. At the higher category level, functions related to the mobilome, including transposase; inorganic ion metabolism, such as iron, cadmium, and copper; defence mechanisms, including restriction-modification, efflux pump, and toxin-antitoxin module; and secretion, such as the type IV secretory pathway, were significantly enriched in the plasmids compared with the chromosomes ($p < 0.05$, Fisher's exact test). In contrast, functions involved in carbohydrate metabolism were significantly higher ($p < 0.05$) in the chromosomes than in the plasmids (Fig. 2.5 and Supplementary materials: Table S2.13).

We further investigated ARGs using the Resfams database (Gibson et al., 2015) and found that a total of 86 plasmids, including four novel plasmid CCs, were positive for ARG-related genes (Supplementary materials: Table S2.14). Many of the hosts were Proteobacteria, accounting for ~76% of the ARG-positive plasmids, Firmicutes with ~20%, and a very few Bacteroidetes, but no plasmid was associated with Actinobacteria (Fig. 2.2a and Supplementary materials: Figure S2.14a). The frequency of ARGs was similar between the plasmids and chromosomes of Proteobacteria and Firmicutes but lower in the plasmids than in the chromosomes of Bacteroidetes (Supplementary materials: Figure S2.14b). A comparison of ARG-positive and ARG-negative plasmids found that ARGs were more frequently encoded by lower-abundance plasmids ($p = 2.1\mathrm{e}{-08}$, Wilcoxon rank-sum test, Fig. 2.6). Overall, the present study found several specific functions more frequently harboured by plasmids than by chromosomes in the IGCJ dataset.

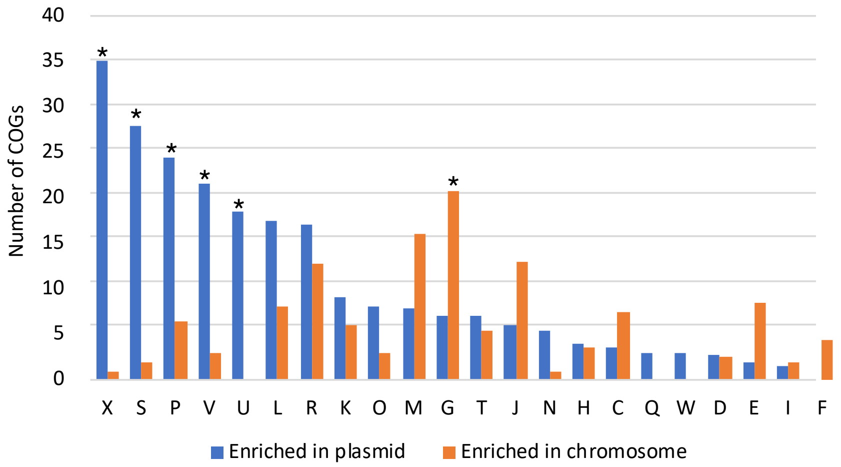

Figure 2.5: Comparison of COG categories between plasmids and chromosomes. The frequency of COGs is compared between 315 relatively high-abundance plasmid clusters and 249 chromosomes (≥0.1% average abundance) in the IGCJ dataset. COG categories with significant differences in enrichment between plasmids and chromosomes are marked with asterisks ($p < 0.05$, Fisher's exact test). Biological functions are abbreviated by letters; X: Mobilome: prophages, transposons; S: function unknown; P: inorganic ion transport and metabolism; V: defence mechanisms; U: intracellular trafficking, secretion, and vesicular transport; L: replication, recombination and repair; R: general function prediction only; K: transcription; O: posttranslational modification, protein turnover, chaperones; M: cell wall/membrane/envelope biogenesis; G: carbohydrate transport and metabolism; T: signal transduction mechanisms; J: translation, ribosomal structure and biogenesis; N: cell motility; H: coenzyme transport and metabolism; C: energy production and conversion; Q: secondary metabolite biogenesis; W: extracellular structures; D: cell cycle control, cell division, chromosome partitioning; E: amino acid transport and metabolism; I: lipid transport and metabolism, transport and catabolism; F: nucleotide transport and metabolism.

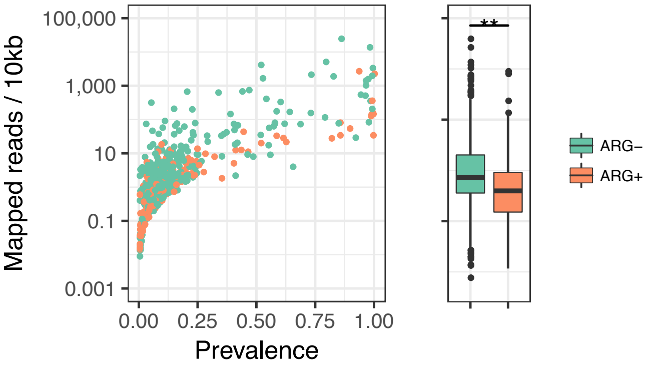

Figure 2.6: Analysis of ARGs in plasmids. The left dot plot shows the prevalence and abundance of 86 ARG-positive (Supplementary materials: Table S2.14) and 229 ARG-negative plasmids according to the Resfams database. The y-axis shows the number of mapped reads per 10-kb region on a log scale. The ARG-positive and ARG-negative plasmids are coloured orange and green, respectively. The right box plot shows the abundance distributions of plasmids with or without ARGs, and their difference is significant as ** denotes $p < 0.01$ (Wilcoxon's rank-sum test).

2.4 Conclusions and Discussion

The present study demonstrated that long-read metagenomic sequencing was useful for the identification of eMGEs as complete contigs and for the exploration of plasmidome entities in the human gut. The plasmid CCs identified by long-read metagenomics included several highly similar but distinct plasmids, which were hard to distinguish by standard short-read metagenomics. This outcome may be the typical case for insufficient assembly of short reads in the metagenomics of communities containing highly similar sequences longer than the read length. The efficient and accurate reconstruction of eMGEs by long-read metagenomics was achieved by two major steps: we first assembled long reads into contigs using the FALCON assembler, which was originally developed for the assembly of diploid genomes with structural variations without dividing contigs, in a more conservative manner (Chin et al., 2016), and then processed the assembled contigs with the output binning results of the contigs (see the "Methods" section). Additionally, a remarkable characteristic of the present approach is its ability to identify relatively high-abundance gut eMGEs independent of their sizes, as demonstrated by the reconstruction of two large plasmid CCs with >600 kb, thereby resulting in the efficient discovery of many novel eMGEs (64/82, 78%).

The 82 CCs were classified as 71 plasmids and 11 phages using several classification assessments (Supplementary materials: Table S2.4). However, one plasmid CC (FA1-2_000589F in Cluster_256) hit a viral contig shorter than the CC, and four similar plasmid CCs in Cluster_461, which were plasmid-positive by PlasFlow and had partial similarity to a known plasmid pBFUK1, hit several viral contigs. Considering the relatively high abundance of these CCs and the lack of typical structural characteristics of prophages in these CCs, these discrepancies could be explained by contamination of non-viral DNA in the VLPs; hence, these CCs are likely to represent plasmids.

The mapping analysis of IGCJ metagenomic reads showed that the ratio of novel eMGEs was ~60%, more than twice the coverage (~20%) of known eMGEs alone (Fig. 2.3b). As described above, because we excluded the plasmids unevenly mapped by non-specific reads from quantification, the observed coverage of the three types of eMGEs may be slightly affected by potential overestimation based on shared genes. The analysis also revealed low coverage of the known plasmid clusters alone, although they represented a large proportion of the plasmid clusters (509/563, 90%). This is probably because they are composed mostly of the plasmids of Proteobacteria species with relatively low abundance in the human gut. In other words, the present study efficiently identified many plasmids hitherto unknown but abundant in the human gut.

It was reported that crAssphages were identified as circular genomes (Dutilh et al., 2014; Guerin et al., 2018). However, our analysis provided evidence suggesting that the five crAssphages had linear genomes with TDRs (Fig. 2.2 and Supplementary materials: Figure S2.5). In a previous study, a circular crAssphage genome was validated by gap closing between fragmented contigs by PCR, followed by sequencing of PCR products (Dutilh et al., 2014). However, PCR amplification between unconnected TDRs in the linear genome is also feasible by duplex formation via annealing between downstream TDRs in the extended DNAs primed from the flanking regions of TDRs, similar to the mechanism for extended primer dimer formation or template switching (Patel et al., 1996), although we cannot exclude the possibility of coexistence of both circular and linear crAssphage genomes.

Although crAssphages were also reported to be highly abundant in the human gut, the ratio of mapped reads varied from 0.03% (JP) to 1.4% (US) among the five countries (Fig. 2.3 and Supplementary materials: Figure S2.12). In addition, the proportion of crAssphage-positive subjects was as low as 60% on average in the 413 individuals (Supplementary materials: Figure S2.12). These data suggest high variability in crAssphages at both the individual and country levels and the presence of two types of gut microbiomes: those with high and low abundance of crAssphages. However, we could not link the abundance and prevalence of crAssphages to the overall microbial composition or the host's genetic background, age, BMI, and sex (Supplementary materials: Figure S2.13). There are several questions that arise from these data. For example, what is the real role of crAssphages in the gut ecosystem? and what is the factor affecting this dominant phage?

The ratio of plasmids to microbial chromosomes in the human gut metagenome has not previously been reported. Our first estimation suggested that plasmids outnumber the microbial cells in IGCJ gut microbiomes. On the other hand, the estimated ratio of crAssphages to microbial cells is approximately consistent with previous estimations of gut phages to microbial cells, ranging from 0.1:1 to 1:1 (Reyes et al., 2012; Kim et al., 2011). The present estimate remains tentative because yet-unidentified eMGEs should exist and will need to be confirmed with more samples.

Host prediction is a challenging issue in eMGE study (Dib et al., 2015; Edwards et al., 2015). A similarity search for the draft genomes of individual cultured species containing unidentified plasmid sequences is a simple but solid method for host assignment of plasmids, once plasmids are identified as complete CCs. Indeed, in this study, hosts for 36 of the 71 plasmid CCs were assigned by a similarity search for draft genomes, of which 13 hosts were also predicted by CO and/or MM to taxonomically close species assigned by the similarity search. In addition, the hosts of two plasmid CCs predicted by both CO and MM and those of four phage CCs predicted by both MM and CRISPR spacer were taxonomically consistent between the two methods (Supplementary materials: Tables S2.9 and S2.10). Thus, there was almost no inconsistency in host prediction between at least two different methods, and many of the predicted hosts were taxonomically assigned at the species and genus levels, demonstrating the practical usefulness of the three methods and their combined use for host prediction of eMGEs, as well as the Hi-C method recently developed (Stewart et al., 2018). In addition, the overall sequence similarity shown here could also be a useful index for host prediction of plasmids, because plasmids from taxonomically similar hosts tended to have relatively high sequence similarities between them (Fig. 2.2a).

In host prediction of phages, YS1-2_2434 and FAKO27_000271F may be novel phages of putative hosts *Bifidobacterium* and *Faecalibacterium*, respectively, because they differed from the recently reported prophages of these two taxa (Duranti et al., 2017; Cornuault et al., 2018). FAKO27_000238F may also be a novel phage and the first associated with *Phascolarctobacterium* as a putative host.

The present analysis also revealed the largest host-plasmid network and the highest abundance of plasmids in Bacteroidetes, which was nearly independent of the overall microbial composition. These results may accord with the previous findings that there was no profound association between the dominant species and its mobile genes and the extensive DNA transfer between *Bacteroidales* species in the human gut (Brito et al., 2016; Coyne et al., 2014). Taken together, our data strongly suggest that Bacteroidetes-associated plasmids are the major players and mediators in modulating human gut microbiome structure and function toward improving the adaptability of the host to environmental changes such as an increase in heavy metal ions.

The functional analysis identified several plasmid-enriched functions, such as transposase, toxin-antitoxin, type IV secretion system (conjugation), and inorganic ion transport (Fig. 2.5 and Supplementary materials: Table S2.13). Among the genes in category X, transposase-related COGs were exclusively identified as plasmid-enriched genes, which may be partly because category X is biased toward many transposases in its composition. While the former three functions were known to be plasmid-enriched (Smillie et al., 2010; Ogilvie et al., 2012), we also found the dissemination of resistance and efflux systems for metal ions such as copper, arsine, tellurium, and cadmium in gut plasmids, suggesting that gut plasmids are determinants of metabolism for toxic metal ions (Silver and Walderhaug, 1992). Our data also revealed that antibiotic functions were strongly linked to relatively low-abundance Proteobacteria plasmids, particularly *Enterobacteriaceae*, in the human gut (Fig. 2.2a and Supplementary materials: Table S2.14), suggesting associations between nosocomial *Enterobacteriaceae* species and the human gut microbiome (San Millan, 2018). However, at present, we do not know the biological significance of the tendency to carry plasmids encoding antibiotic functions more frequently in low-abundance species than high-abundance plasmids.

In conclusion, long-read metagenomics provides an efficient method for the exploration of uncharted eMGEs in the human gut, and the accumulated data represent an alternative resource useful for a deeper understanding of human gut microbial ecology.

2.5 Supplementary materials

2.5.1 Supplementary Figures

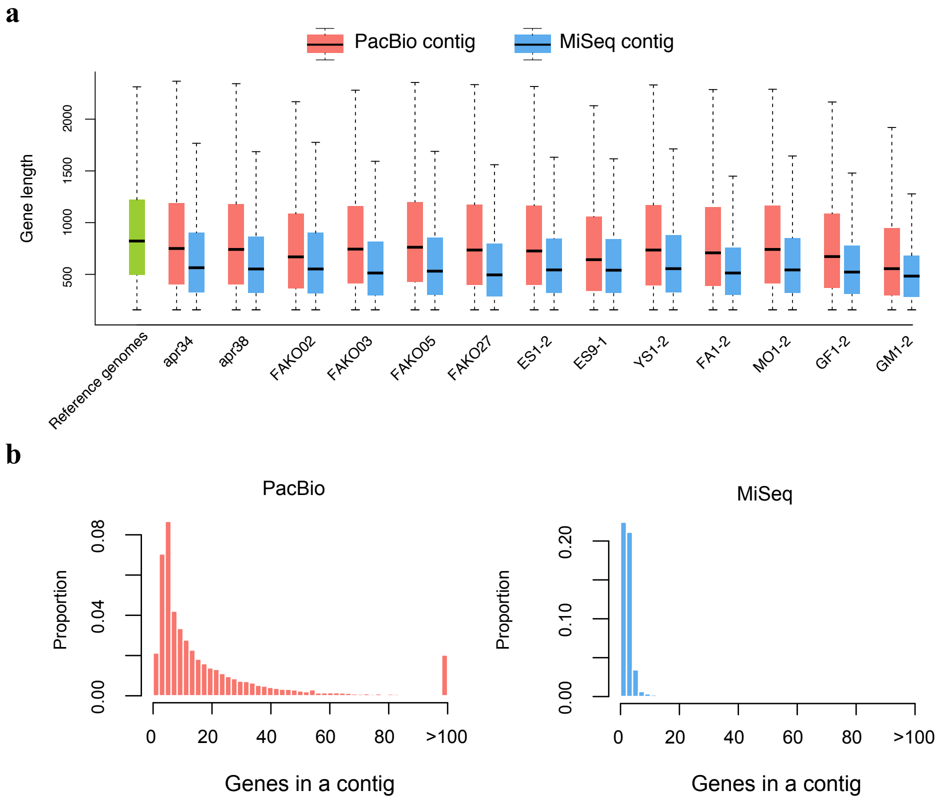

Figure S2.1: Genes in PacBio and short-read contigs. **a** Comparison of length distributions of genes identified in PacBio and short-read (MiSeq) contigs and reference genomes containing complete genomes. The box plots show IQR by boxes, medians by central lines, and the lowest and highest values within 1.5 times the IQR are shown by whiskers. For visualization, outliers are not shown in this figure. **b** Histograms for the number of genes identified in the PacBio and MiSeq contigs.

a

Identity	Alignment coverage	E-value	Sensitivity	Specificity
-	20%	1e-5	97%	82%
-	30%	1e-5	97%	84%
-	40%	1e-5	97%	87%
-	50%	1e-5	97%	90%
-	60%	1e-5	96%	91%
-	70%	1e-5	96%	92%
-	80%	1e-5	96%	93%
-	90%	1e-5	95%	95%
20%	-	1e-5	97%	82%
30%	-	1e-5	97%	83%
40%	-	1e-5	95%	92%
50%	-	1e-5	90%	95%
60%	-	1e-5	85%	97%
70%	-	1e-5	80%	98%
80%	-	1e-5	76%	99%
90%	-	1e-5	73%	99%
20%	20%	1e-5	97%	82%
30%	30%	1e-5	97%	85%
40%	40%	1e-5	94%	94%
50%	50%	1e-5	89%	97%
60%	60%	1e-5	83%	98%
70%	70%	1e-5	79%	99%
80%	80%	1e-5	75%	99%
90%	90%	1e-5	72%	99%

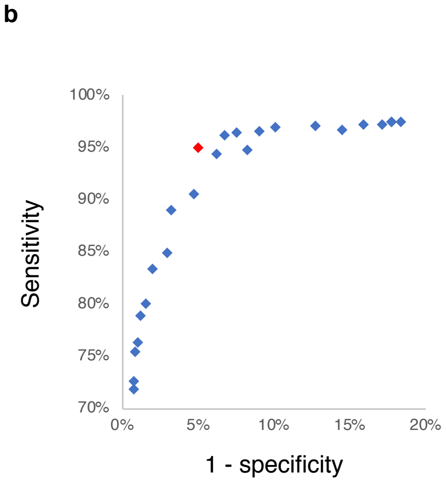

Figure S2.2: Optimization for identification of POGs. **a**, Estimation of sensitivities (the number of phages from which POG(s) were detected / the number of phages) and specificities (the number of non-phages from which POG(s) were not detected / the number of non-phages) with various thresholds. Calculations were performed by aligning all predicted genes of the reference plasmid and phage sequences to POGs. **b** Relation between the sensitivity (y-axis) and the false positive ratio (1 − specificity). A red dot is the nearest to the perfect prediction at the upper left corner (100% sensitivity and 100% specificity) among the thresholds tested under the conditions of alignment coverage ≥90% without a threshold for identity.

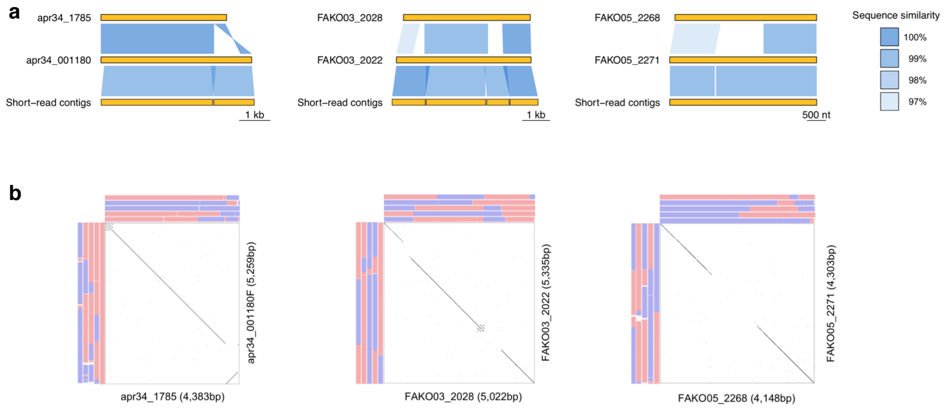

Figure S2.3: Sequence alignments of two highly similar but distinct plasmid CCs in three samples. **a** Alignment of three pairs of similar plasmid CCs identified in three subjects. Orange bars represent two highly similar plasmid CCs (upper two) generated from PacBio reads and the corresponding short-read contigs (bottom) in each sample. Multiple fragmented short-read contigs (left and middle) were aligned with the plasmid CCs, and a short-read contig (right) was aligned with one of either plasmid CCs. The similar regions are connected with blue rectangles, of which shades indicate the degree of sequence similarity between them. **b** Dot plots of two similar plasmid CCs in three samples. PacBio subreads covering the forward and reverse strands of the entire CCs are shown by red and blue bars, respectively. Only a part of the mapped subreads is shown here.

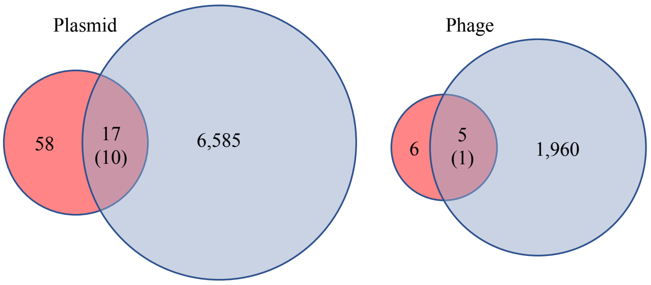

Figure S2.4: Similarity search of 82 CCs against the public plasmid/phage database. The Venn diagrams show 71 plasmid and 11 phage CCs (red) identified in this study and known plasmids and phages in GenBank (blue), respectively. The 17 plasmid CCs and five phage CCs were matched with 10 known plasmids and one phage (crAssphage) in GenBank, respectively

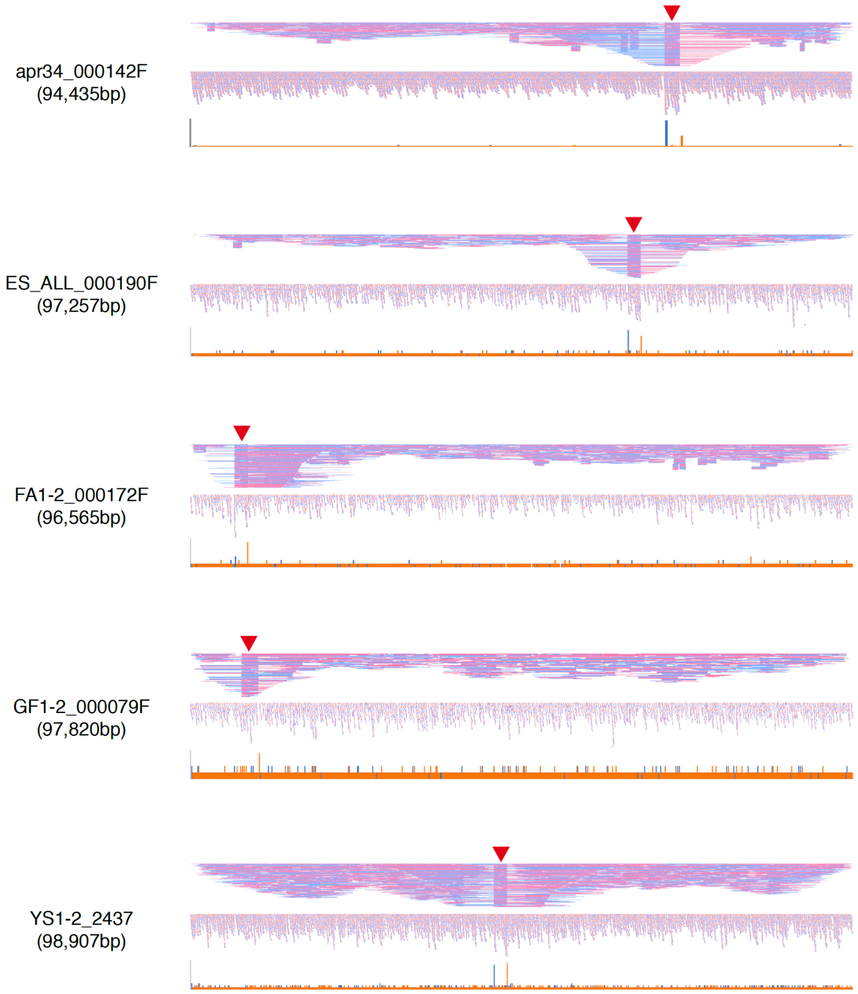

Figure S2.5: Mapping of PacBio subreads and short reads to the five crAssphage CCs. Alignments of PacBio long subreads and MiSeq short reads to the five crAssphage CCs are shown in the upper and middle diagrams, respectively. Red and blue horizontal lines represent forward and reverse reads mapped to the CCs, respectively. Red inverse triangles highlight the region of TDRs with approximately two times higher number of mapped reads than others in the CCs. Alignments with excessive PacBio subreads were eliminated from the diagrams. The bottom diagram shows the frequency of start sites (5' position) of aligned short reads in the CCs. The orange and blue bars represent the numbers of forward and reverse reads, respectively.

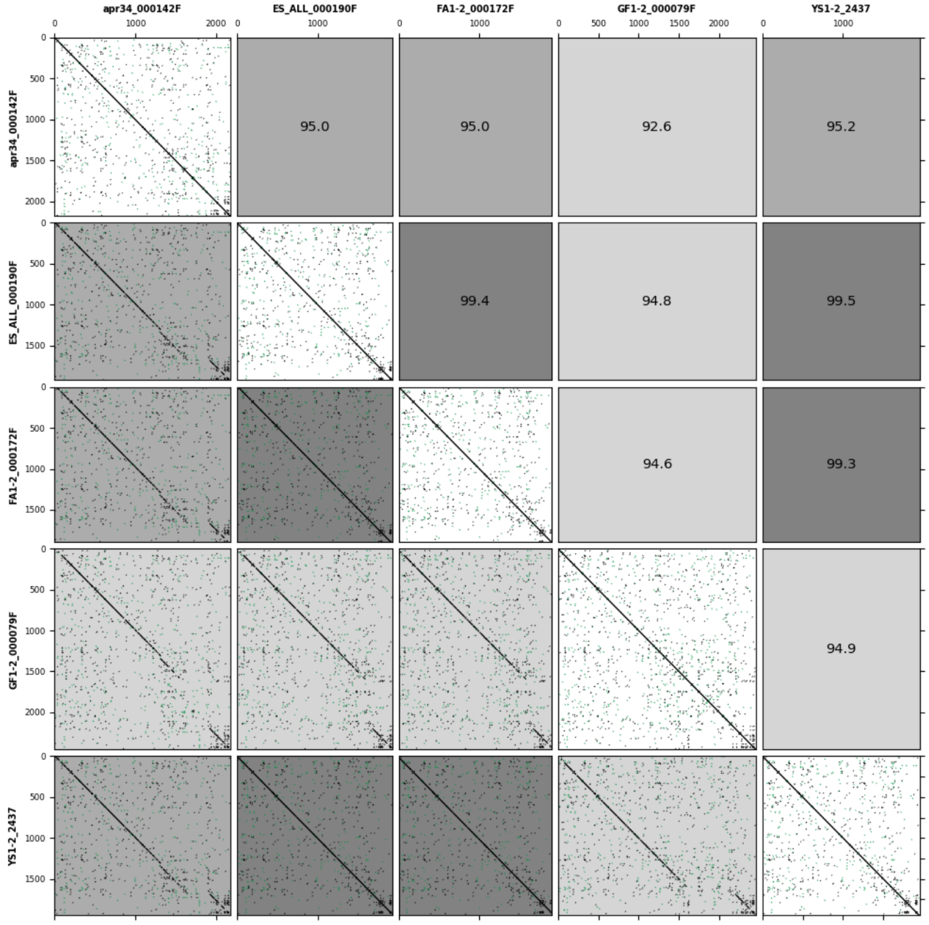

Figure S2.6: Dot plot of TDRs in the five crAssphages. Dot plots of all pairs of TDRs in the five crAssphage genomes are shown. The numbers in the matrix denote percentage identities between the two TDRs.

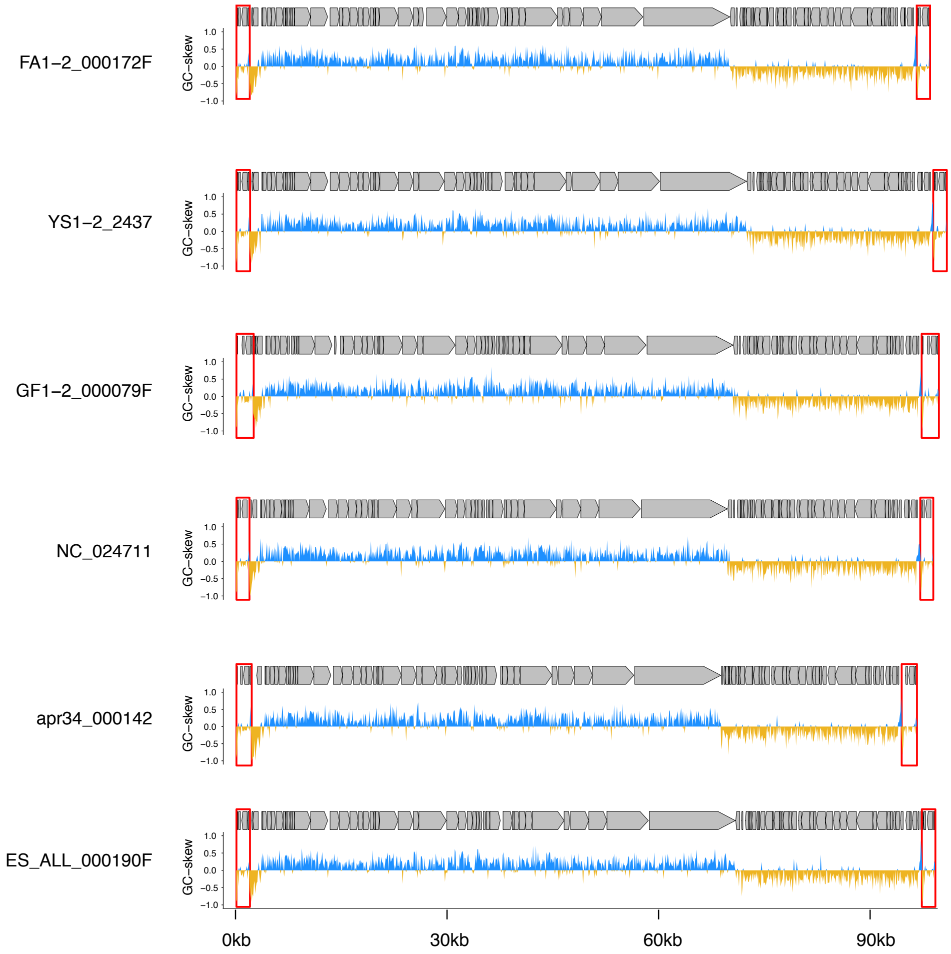

Figure S2.7: GC skews in the linear crAssphage genomes. Grey pentagons indicate putative genes in the crAssphage genomes. TDRs are indicated by red boxes. GC skews are shown in blue and orange.

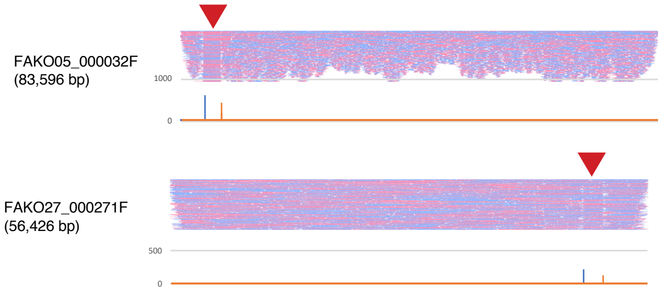

Figure S2.8: Mapping of PacBio subreads to two phage CCs. Alignments of PacBio subreads mapped to two phage CCs (FAKO05_000032F and FAKO27_0000271F) are shown as described in Fig. S2.5. The data suggest that these two phage CCs have linear genomes.

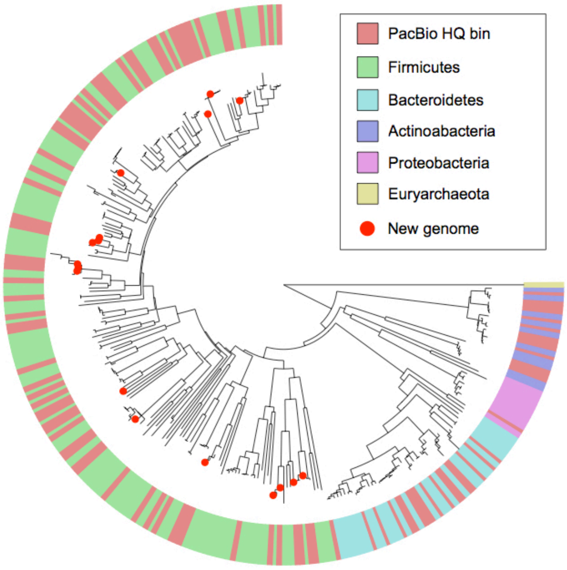

Figure S2.9: Phylogenetic tree of 101 HQ chromosome bins and 181 known genomes. A neighbour-joining phylogenetic tree was constructed from 101 HQ genome bins and 181 known genomes of four phyla in GenBank with Euryarchaeota (Methanobrevibacter smithii) as an outgroup. Five phyla are shown in different colours (green for Firmicutes, purple for Actinobacteria, pink for Proteobacteria, blue for Bacteroidetes, and yellow for Euryarchaeota), and red for 101 HQ genome bins in the outer circle. Red circles on the tree edges indicate 17 novel genomes phylogenetically distinct from the known genomes.

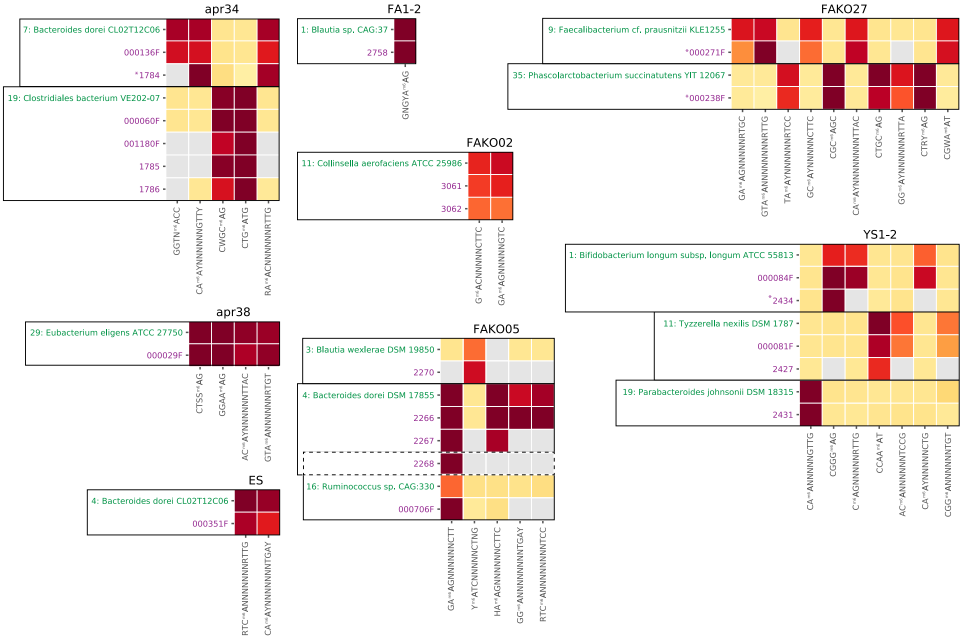

Figure S2.10: Host prediction by MM similarity between eMGEs and HQ chromosome bins in the PacBio JP dataset. The eMGEs (purple) and host strains (green) having common MMs in eight subjects are shown with separate boxes. The eMGEs marked with asterisks indicate phages, and others are plasmids. The MMs with m6A are shown at the bottom of each box. Red shades indicate mean IPD ratio values higher than the threshold of 2.5, and yellow indicates mean IPD ratios less than the threshold. Grey denotes the absence of common MMs between host strains and eMGEs. The eMGE 2268 links with two different host strains are boxed by a dashed line.

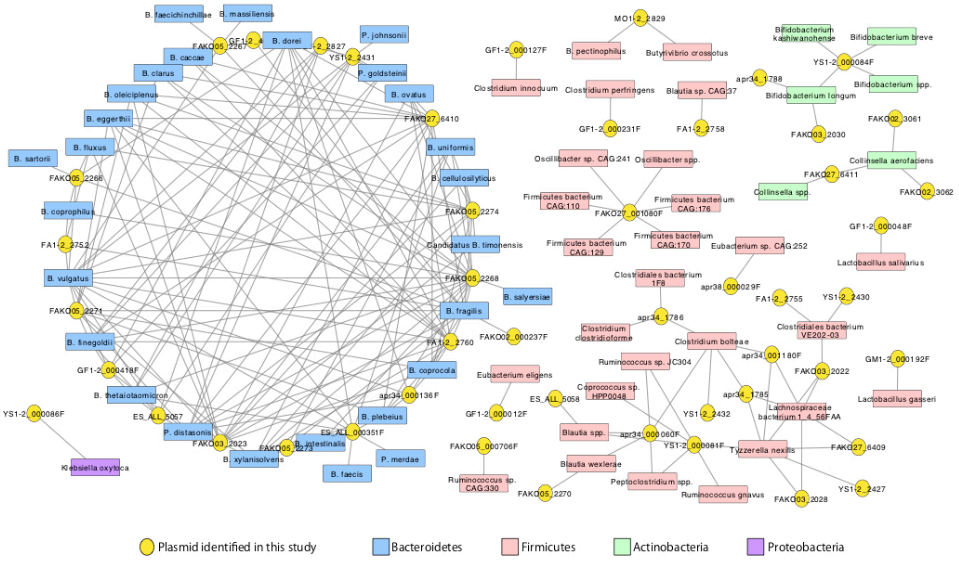

Figure S2.11: Host-plasmid network. The predicted host-plasmid relationships were summarized and visualized as a network. The circles and squares show plasmid CCs identified in this study and predicted hosts, respectively. The colours of the squares indicate host taxonomy at the phylum level (pink for Firmicutes, green for Actinobacteria, purple for Proteobacteria, blue for Bacteroidetes).

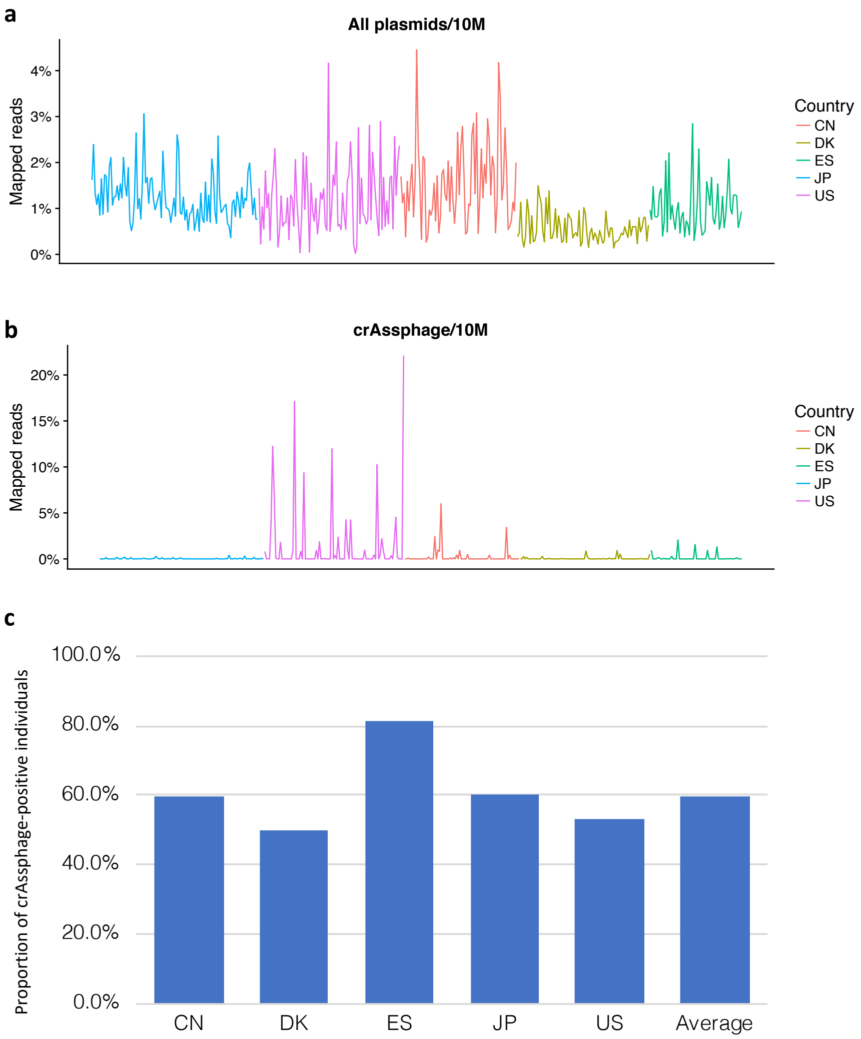

Figure S2.12: Ratios of reads mapped to plasmids and crAssphages in 413 metagenomic data sets and proportions of crAssphage-positive individuals. **a-b** Metagenomic reads from 413 individuals (10 M reads per individual) in the IGCJ dataset are mapped to eMGE and crAssphage clusters. The x-axis shows 413 individuals from China (orange), Denmark (brown), Spain (green), Japan (light blue), and the US (pink). The y-axis shows the ratio of reads mapped to the clusters. **c** Proportions of crAssphage-positive individuals (≥1 mapped read) in the five countries.

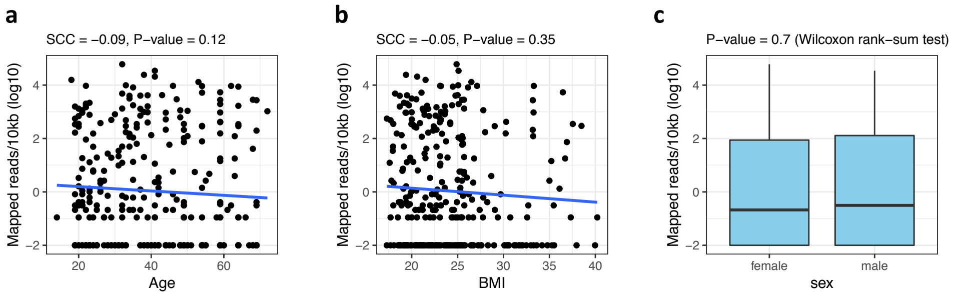

Figure S2.13: Association analysis of the abundance of crAssphages with subjects' age, BMI and sex in the IGCJ dataset. SCCs between crAssphage abundance and age (**a**) and BMI (**b**). Each circle represents each subject, and the blue line is the regression line. Comparison of crAssphage abundance between male and female participants (**c**). Pseudo-count (0.00001) was added to the abundance, and the values were log-transformed. The publicly available metadata (age, BMI, and sex) of 323 subjects in four countries (except the US subjects) were used for the analysis.

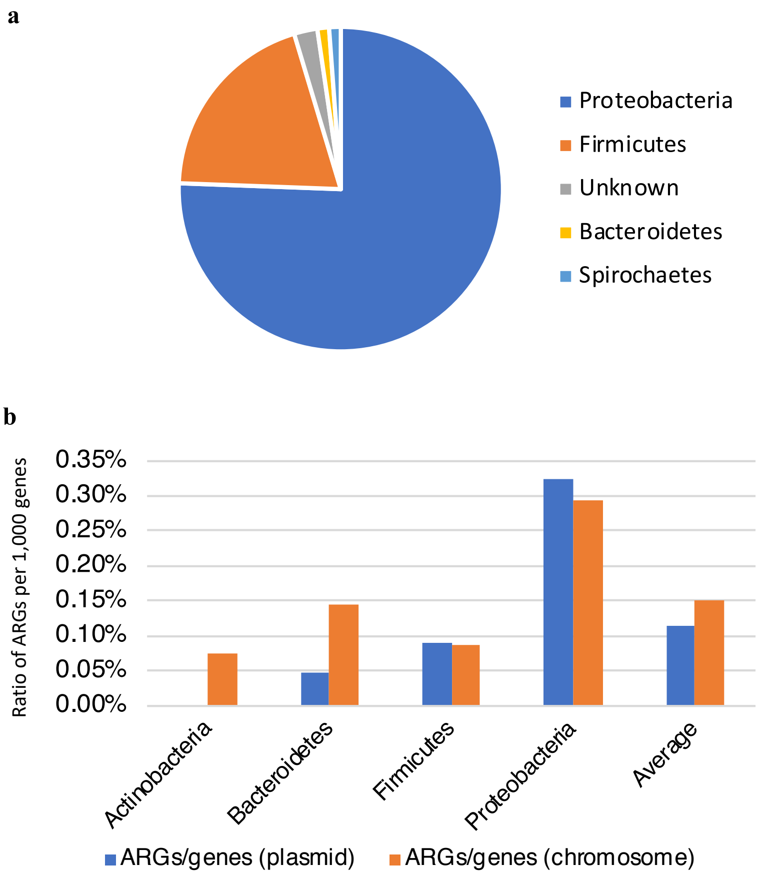

Figure S2.14: ARGs in plasmids in the IGCJ dataset. **a** Proportion of host phyla of plasmids containing ARGs based on the Resfams database. **b** Ratio of ARGs per 1,000 genes in plasmids and chromosomes is shown.

2.5.2 Supplementary Tables

Table S2.1: Summary of metagenomic sequencing of fecal samples from 12 individuals by PacBio and other sequencers.

Subject ID	Age	Sex	Family	PacBio RS II (P4-C2)			PacBio RS II (P6-C4)			Total # subreads (PacBio)	Total base pairs (PacBio)	Total base pairs (PacBio, error-corrected reads)
				# Subreads	Average subread length	Base pairs	# Subreads	Average subread length	Base pairs			
apr34	20	female	no	-	-	-	1,914,637	4,892	9,365,631,339	1,914,637	9,365,631,339	4,705,237,154
apr38	19	male	no	-	-	-	870,751	8,665	7,545,057,362	870,751	7,545,057,362	5,216,585,842
FAKO02	47	male	no	-	-	-	994,536	9,862	9,807,837,159	994,536	9,807,837,159	6,178,480,285
FAKO03	50	male	no	1,649,889	1,911	3,152,770,640	-	-	-	1,649,889	3,152,770,640	199,064,009
FAKO05	50	male	no	1,594,348	2,294	3,657,589,101	-	-	-	1,594,348	3,657,589,101	433,924,565
FAKO27	50	male	no	3,335,131	2,573	8,580,950,550	1,967,478	4,488	8,829,849,357	5,302,609	17,410,799,907	4,571,886,382
ES1-2	17	female	yes	-	-	-	1,511,486	7,715	11,661,252,096	1,511,486	11,661,252,096	-
ES9-1	17	female	yes	-	-	-	1,334,470	8,024	10,707,762,806	1,334,470	10,707,762,806	-
ES_ALL (ES1-2 + ES9-1)	17	female	yes	-	-	-	2,845,956	7,860	22,369,014,902	2,845,956	22,369,014,902	8,526,388,837
FA1-2	44	male	yes	-	-	-	1,708,252	7,000	11,957,772,759	1,708,252	11,957,772,759	3,367,414,284
GF1-2	77	male	yes	-	-	-	1,744,355	7,954	13,874,353,103	1,744,355	13,874,353,103	3,535,433,781
GM1-2	78	female	yes	-	-	-	1,175,021	9,289	10,914,987,891	1,175,021	10,914,987,891	2,109,877,682
MO1-2	44	female	yes	-	-	-	1,526,066	7,195	10,979,508,573	1,526,066	10,979,508,573	1,741,880,643
YS1-2	5	female	yes	-	-	-	1,820,072	7,643	13,910,897,854	1,820,072	13,910,897,854	5,313,556,355
Avearge of 13 samples										1,780,499	10,380,478,507	3,397,576,453
Average of 11 subjects (except FAKO03 & FAKO05)							1,506,102	7,216	10,868,628,209	1,990,226	12,813,586,085	4,526,674,125

Subject ID	Total base pairs (PacBio, mapped to hg19)	Total # subreads (PacBio, mapped to hg19)	Roche 454		Ion PGM		Illumina MiSeq		Total reads	Total base pairs	% mapped reads to human genome (MiSeq)
			Reads	Base pairs	Reads	Base pairs	Reads	Base pairs	(short read)	(short read)	
apr34	51,870	20 (0.001%)	1,041,477	679,049,047	-	-	17,913,411	5,260,094,469	18,954,888	5,939,143,516	-
apr38	185,488	28 (0.003%)	903,832	552,619,941	-	-	19,238,817	5,704,934,670	20,142,649	6,257,554,611	-
FAKO02	147,275	15 (0.002%)	1,212,255	774,334,821	4,977,020	1,191,843,698	20,601,066	3,171,189,959	26,790,341	5,137,368,478	-
FAKO03	40,685	19 (0.001%)	813,005	474,528,076	3,060,941	722,273,852	23,416,757	3,635,198,059	27,290,703	4,831,999,987	-
FAKO05	28,848	12 (0.001%)	1,013,519	772,877,228	4,135,434	983,799,832	18,049,491	2,806,997,676	23,198,444	4,563,674,736	-
FAKO27	81,768	30 (0.001%)	1,116,983	876,718,876	6,178,406	1,358,286,547	25,430,117	3,958,248,462	32,725,506	6,193,253,885	-
ES1-2	2,736	1 (0.000%)	-	-	-	-	12,351,693	-	12,351,693	2,963,368,976	0.001%
ES9-1	293,536	33 (0.002%)	-	-	-	-	10,009,420	-	10,009,420	2,341,061,545	0.006%
ES_ALL (ES1-2 + ES9-1)	-	-	-	-	-	-	22,361,113	-	22,361,113	5,304,430,521	-
FA1-2	485,320	68 (0.004%)	-	-	-	-	3,491,891	-	3,491,891	722,718,336	0.059%
GF1-2	12,069	1 (0.000%)	-	-	-	-	13,721,388	-	13,721,388	2,957,324,378	0.002%
GM1-2	11,735	1 (0.000%)	-	-	-	-	3,793,400	-	3,793,400	712,273,902	0.009%
MO1-2	6,491	2 (0.000%)	-	-	-	-	11,685,661	-	11,685,661	2,654,297,840	0.003%
YS1-2	0	0 (0%)	-	-	-	-	11,140,944	-	11,140,944	2,433,077,447	0.002%
Avearge of 13 samples									17,941,411	3,975,593,136	
Average of 11 subjects (except FAKO03 & FAKO05)											

Table S2.2: Contigs generated from assembly of PacBio and short reads. Note that contigs shorter than 500 bp were removed in the short-read assemblies.

Subject ID	PacBio contig#	PacBio contig length in bp	PacBio contig N50 length in bp	PacBio longest contig length in bp	PacBio shortest contig length in bp	Short-read contig#	Short-read contig length in bp	Short-read contig N50 length in bp	Short-read longest contig length in bp	Short-read shortest contig length in bp
apr34	1,887	99,177,571	255,972	2,976,383	95	85,631	177,020,327	4,395	236,789	500
apr38	2,048	101,254,602	279,204	4,227,722	101	111,996	212,659,740	3,259	238,535	500
FAKO02	3,247	141,555,390	139,859	2,723,705	98	84,000	146,324,073	4,599	369,949	500
FAKO03	1,850	18,552,324	24,641	467,305	65	45,290	59,492,962	3,286	293,866	500
FAKO05	2,132	28,542,572	74,841	2,402,855	78	74,229	168,779,741	4,510	345,087	500
FAKO27	6,277	111,069,464	65,539	2,034,231	58	100,546	185,013,815	2,981	320,160	500
ES_ALL	5,153	138,117,496	115,586	2,371,174	130	84,559	182,493,422	3,006	277,431	500
FA1-2	2,840	96,205,212	129,414	1,206,678	142	117,994	217,785,539	1,595	65,052	500
GF1-2	3,988	83,619,483	68,042	5,332,272	78	107,637	147,829,304	1,669	250,990	500
GM1-2	4,165	61,767,800	30,309	814,881	128	49,767	51,891,145	1,091	173,512	500
MO1-2	2,807	60,994,291	67,338	1,019,902	74	77,575	133,388,428	3,019	243,440	500
YS1-2	2,542	102,616,022	151,814	2,253,333	108	51,443	96,587,603	3,282	217,680	500
Total	38,936	1,043,472,227				990,667	1,779,266,099			

Table S2.3: Distribution of read depths and contig lengths in PacBio contigs of the three subjects.

Read depth	Contig length in bp	Ratio (%)
<5	278,600	0.16
≥5	3,600,101	2.10
≥10	10,115,369	5.89
≥15	13,894,927	8.09
≥20	19,726,799	11.49
≥25	17,286,297	10.07
≥30	20,226,259	11.78
≥35	10,026,977	5.84
≥40	5,897,292	3.44
≥45	5,736,580	3.34
≥50	7,274,777	4.24
≥55	3,272,158	1.91
≥60	4,564,614	2.66
≥65	10,843,639	6.32
≥70	7,525,208	4.38
≥75	10,288,008	5.99
≥80	5,075,999	2.96
≥85	2,615,731	1.52
≥90	6,415,924	3.74
≥95	3,812,539	2.22
≥100	3,197,846	1.86

Table S2.4: Classification and characterization of 82 CCs <1-Mb in PacBio assembly. *Results of phage classification are shown by 1 for "most confident", 2 for "likely", 3 for "possible", and - for "no score". **1 and 0 indicate presence and absence of plasmid-enriched genes such as plasmid replication, toxin-antitoxin system, mobilization protein and type IV secretion system in CCs, respectively.

Name	Length	Classification				Plasmid-enriched	Novelty	High-similar	High-similar references in databases	Identity with closest reference	Alignment	Contig length / Reference length
	in bp	POG*	Virsorter*	PlasFlow	This study	genes**		reference ID			coverage	
apr34_000060F	243,670	-	-	Chromosome	Plasmid	COG3451, COG3505, COG3077, COG3077, ParA (prokka)	Novel	-	-	-	-	-
apr34_000136F	98,299	-	-	Chromosome	Plasmid	COG3451, COG3843, COG3843, COG2336, ParA (prokka)	Novel	-	-	-	-	-
apr34_000142F	94,435	1	2	Chromosome	Phage	-		NC_024711.1	Uncultured phage crAssphage	93.2%	92.8%	0.97
									phage: 3300014961_Ga0134526_1000359	96.4%	23.7%	3.25
									phage: 3300006477_Ga0100232_103270	95.2%	5.6%	15.43
									phage: 3300014804_Ga0134371_1002626	98.4%	5.6%	18.03
apr34_001180F	5,259	-	-	Unclassified	Plasmid	0	Novel	-	-	-	-	-
apr34_1784	45,225	1	1	Chromosome	Phage	-	Novel	-	-	-	-	-
apr34_1785	4,383	-	-	Unclassified	Plasmid	COG3843	Novel	-	-	-	-	-
apr34_1786	5,602	-	-	Plasmid	Plasmid	0	Novel	-	-	-	-	-
apr34_1788	165,458	-	-	Chromosome	Plasmid	0	Novel	-	-	-	-	-
apr34_1792	30,579	1	2	Chromosome	Phage	-	Novel	-	-	-	-	-
apr38_000029F	617,950	-	-	Chromosome	Plasmid	COG1475	Novel	-	-	-	-	-
apr38_000077F	216,202	-	-	Chromosome	Plasmid	COG3077, COG3077, COG3505, COG3451, ParA (prokka)	Novel	-	-	-	-	-
apr38_2079	28,607	-	-	Chromosome	Plasmid	0	Novel	-	-	-	-	-
apr38_2081	47,897	-	-	Unclassified	Plasmid	0	Novel	-	-	-	-	-

Name	Length	Classification				Plasmid-enriched	Novelity	High-similar	High-similar references in databases	Identity with closest reference	Alignment	Contig length / Reference length
	in bp	POG*	Virsorter*	PlasFlow	This study	genes**		reference ID			coverage	
ES_ALL_000190F	97,257	1	2	Chromosome	Phage	-		NC_024711.1	Uncultured phage crAssphage	97.1%	94.7%	1.00
									phage: 3300008360_Ga0114875_1003073	97.9%	9.0%	8.72
									phage: 7000000075_SRS016056_WUGC_scaffold_1507	96.3%	10.3%	9.50
									phage: 3300009692_Ga0116171_10015780	98.8%	5.9%	16.86
									phage: 3300009711_Ga0116166_1006214	98.5%	8.4%	11.91
									phage: 3300009687_Ga0116144_10011271	98.1%	7.1%	13.99
									phage: 3300009687_Ga0116144_10012116	98.7%	6.8%	14.72
ES_ALL_000351F	44,025	-	-	Unclassified	Plasmid	COG3505, COG3843, COG5527, COG3451, ParA (prokka)		NC_006873.1	Bacteroides fragilis NCTC 9343 pBF9343 plasmid	95.6%	96.1%	1.20
ES_ALL_5057	31,059	-	-	Plasmid	Plasmid	COG3451, COG3505, COG5527		NC_004703.1	Bacteroides thetaiotaomicron VPI-5482 plasmid p5482	98.8%	91.7%	0.94
ES_ALL_5058	65,724	-	-	Chromosome	Plasmid	0	Novel	-	-	-	-	-
ES_ALL_5059	8,278	-	-	Plasmid	Plasmid	COG3077, COG2161, COG3843	Novel	-	-	-	-	-
FA1-2_000172F	96,565	1	2	Chromosome	Phage	-		NC_024711.1	Uncultured phage crAssphage	97.0%	94.5%	0.99
									phage: 3300009687_Ga0116144_10011271	98.1%	7.2%	13.89
									phage: 3300012956_Ga0154020_10017436	98.2%	8.3%	12.01
									phage: 3300009692_Ga0116171_10015780	98.9%	6.0%	16.74
									phage: 7000000075_SRS016056_WUGC_scaffold_1507	96.4%	10.4%	9.43
									phage: 7000000613_SRS023847_Baylor_scaffold_764	98.3%	7.7%	13.07
FA1-2_000250F	62,025	-	-	Chromosome	Plasmid	COG3843, COG3451, COG3505, ParA (prokka)	Novel	-	-	-	-	-
FA1-2_000589F	26,749	-	-	Chromosome	Plasmid	0	Novel	-	phage: 2051223001_mc15b_MC15B_contig06336, IMVGR	97.3%	21.9%	4.26
FA1-2_2752	5,396	-	-	Chromosome	Plasmid	COG2026, COG3843, COG5527	Novel	-	-	-	-	-
FA1-2_2754	6,124	-	-	Plasmid	Plasmid	COG5527	Novel	-	-	-	-	-
FA1-2_2755	10,436	-	-	Unclassified	Plasmid	COG3077	Novel	-	-	-	-	-
FA1-2_2756	50,544	-	3	Chromosome	Plasmid	COG1475, COG3505	Novel	-	-	-	-	-
FA1-2_2758	8,177	-	-	Unclassified	Plasmid	0	Novel	-	-	-	-	-

Name	Length in bp	Classification				Plasmid-enriched genes**	Novelity	High-similar reference ID	High-similar references in databases	Identity with closest reference	Alignment coverage	Contig length / Reference length
		POG*	Virsorter*	PlasFlow	This study							
FA1-2_2760	4,306	-	-	Plasmid	Plasmid	COG5527, COG3843	Novel	NC_019534.1	Bacteroides fragilis plasmid pBFUK1 (partial)	99.9%	100.0%	0.34
								-	phage: Bacteroidia_gi_329959038_4306_bp_DNA_circular_ENV	100.0%	100.0%	1.00
FAKO02_000237F	74,873	-	-	Chromosome	Plasmid	COG3451, COG1475, COG5527, COG3505	Novel	-	-	-	-	-
FAKO02_3061	25,023	-	-	Unclassified	Plasmid	ParA (prokka)	Novel	-	-	-	-	-
FAKO02_3062	54,132	-	-	Chromosome	Plasmid	COG3451, COG3505, ParA (prokka)	Novel	-	-	-	-	-
FAKO02_3063	70,284	-	3	Unclassified	Plasmid	COG3451, COG3451, COG1475, COG3505	Novel	-	-	-	-	-
FAKO03_2022	5,335	-	-	Plasmid	Plasmid	0	Novel	-	-	-	-	-
FAKO03_2023	2,782	-	-	Plasmid	Plasmid	0		NC_005026.1	Bacteroides fragilis IB143 plasmid pBI143	90.1%	98.8%	1.01
FAKO03_2024	6,599	-	-	Plasmid	Plasmid	COG2026	Novel	-	-	-	-	-
FAKO03_2027	11,603	-	-	Unclassified	Plasmid	ParA (prokka)	Novel	-	-	-	-	-
FAKO03_2028	5,022	-	-	Plasmid	Plasmid	COG3843	Novel	-	-	-	-	-
FAKO03_2030	3,680	-	-	Unclassified	Plasmid	COG5527		NC_010861.1	Bifidobacterium longum plasmid p6043B	100.0%	100.0%	1.00
FAKO05_000032F	83,596	1	2	Chromosome	Phage	-	Novel	-	phage: 7000000482_SRS016541_Baylor_scaffold_7470	98.8%	10.1%	8.04
									phage: 3300014833_Ga0119870_1001739	98.4%	10.8%	9.28
									phage: 3300009694_Ga0116170_10004631	98.7%	14.5%	6.68
									phage: 3300009692_Ga0116171_10010906	98.6%	8.7%	11.44
									phage: 3300009692_Ga0116171_10006832	99.0%	11.7%	8.53
FAKO05_000706F	6,322	-	-	Unclassified	Plasmid	0	Novel	-	-	-	-	-
FAKO05_2266	4,970	-	-	Plasmid	Plasmid	COG5527, COG3843, COG2026	Novel	-	-	-	-	-
FAKO05_2267	4,787	-	-	Plasmid	Plasmid	COG5527, COG3843	Novel	-	-	-	-	-
FAKO05_2268	4,148	-	-	Plasmid	Plasmid	COG5527, COG3843	Novel	NC_019534.1	Bacteroides fragilis plasmid pBFUK1 (partial)	98.1%	71.6%	0.32
									phage: Bacteroidia_gi_298484481_4148_bp_DNA_circular_ENV	100.0%	91.4%	1.00
FAKO05_2270	3,842	-	-	Unclassified	Plasmid	0	Novel	-	-	-	-	-
FAKO05_2271	4,303	-	-	Plasmid	Plasmid	COG5527, COG3843	Novel	NC_019534.1	Bacteroides fragilis plasmid pBFUK1 (partial)	100.0%	100.0%	0.34
									phage: Bacteroidia_gi_329959038_4306_bp_DNA_circular_ENV	99.9%	92.9%	1.00
FAKO05_2273	5,594	-	-	Chromosome	Plasmid	COG2336		NC_011073.1	Bacteroides fragilis plasmid pBFP35	99.8%	100.0%	1.00
FAKO05_2274	2,784	-	-	Plasmid	Plasmid	0		NC_005026.1	Bacteroides fragilis IB143 plasmid pBI143	90.2%	99.9%	1.01

Name	Length	Classification				Plasmid-enriched	Novelity	High-similar	High-similar references in databases	Identity with closest reference	Alignment	Contig length / Reference length
	in bp	POG*	Virsorter*	PlasFlow	This study	genes**		reference ID			coverage	
FAKO27_000238F	64,223	1	-	Chromosome	Phage	-	Novel	-	-	-	-	-
FAKO27_000271F	56,426	1	2	Chromosome	Phage	-	Novel	-	-	-	-	-
FAKO27_001080F	12,395	-	-	Unclassified	Plasmid	COG2161	Novel	-	-	-	-	-
FAKO27_6405	2,873	-	-	Unclassified	Plasmid	COG5527	Novel	-	-	-	-	-
FAKO27_6407	2,970	-	-	Unclassified	Plasmid	0	Novel	-	-	-	-	-
FAKO27_6409	5,023	-	-	Plasmid	Plasmid	COG3843	Novel	-	-	-	-	-
FAKO27_6410	4,307	-	-	Plasmid	Plasmid	COG5527, COG3843	Novel	NC_019534.1	Bacteroides fragilis plasmid pBFUK1 (partial)	99.9%	100.1%	0.34
									phage: Bacteroidia_gi_329959038_4306_bp_DNA_circular_ENV	100.0%	95.3%	1.00
FAKO27_6411	44,683	-	-	Unclassified	Plasmid	COG3077	Novel	-	-	-	-	-
FAKO27_6412	7,241	-	-	Plasmid	Plasmid	COG3077	Novel	-	-	-	-	-
FAKO27_6413	2,897	-	-	Unclassified	Plasmid	COG5527, COG2026	Novel	-	-	-	-	-
GF1-2_000012F	666,740	-	-	Chromosome	Plasmid	COG2336, COG1475, COG3077	Novel	-	-	-	-	-
GF1-2_000048F	167,388	-	-	Chromosome	Plasmid	COG5527, ParA (prokka)		NZ_CP011404.1	Lactobacillus salivarius str. Ren plasmid pR1	97.3%	78.4%	0.95
GF1-2_000079F	97,820	1	2	Chromosome	Phage	-		NC_024711.1	Uncultured phage crAssphage	95.7%	92.7%	1.01
									phage: 3300009655_Ga0116190_1011653	98.6%	4.9%	19.52
									phage: 7000000441_SRS015794_Baylor_scaffold_7923	98.4%	11.0%	8.25
									phage: 3300009682_Ga0116172_10005577	97.3%	10.2%	9.81
									phage: 7000000613_C192949	94.5%	8.3%	12.13
									phage: 3300006258_Ga0099398_100973	97.2%	16.9%	5.58
									phage: 3300008482_Ga0115187_100719	96.2%	22.6%	3.50
									phage: 3300009685_Ga0116142_10014943	98.4%	5.7%	17.69
GF1-2_000127F	72,273	-	3	Unclassified	Plasmid	COG3843, COG3451, COG3505, COG5527	Novel	-	-	-	-	-
GF1-2_000231F	48,134	-	-	Plasmid	Plasmid	0		NZ_CP009558.1	Clostridium perfringens strain FORC_003 plasmid pFORC3	100.0%	100.0%	0.85
GF1-2_000286F	37,242	-	-	Unclassified	Plasmid	COG2161, COG3077, COG2336	Novel	-	-	-	-	-
GF1-2_000418F	30,976	-	-	Plasmid	Plasmid	COG5527, COG3505, COG3451		NC_004703.1	Bacteroides thetaiotaomicron VPI-5482 plasmid p5482	99.7%	91.7%	0.94
GF1-2_000511F	26,269	-	-	Plasmid	Plasmid	COG3077, COG3077, COG2161, COG3505	Novel	-	-	-	-	-

Name	Length in bp	Classification				Plasmid-enriched genes**	Novelity	High-similar reference ID	High-similar references in databases	Identity with closest reference	Alignment coverage	Contig length / Reference length
		POG*	Virsorter*	PlasFlow	This study							
GF1-2_4069	117,808	-	3	Unclassified	Plasmid	COG3505, COG2161, COG3451	Novel	-	-	-	-	-
GF1-2_4070	7,660	-	-	Plasmid	Plasmid	COG5527, COG3505, COG5527	Novel	-	-	-	-	-
GM1-2_000192F	44,253	-	-	Unclassified	Plasmid	COG3077, COG2336, COG2161, COG3077		NZ_CP006810.1	Lactobacillus gasseri 130918 plasmid	98.2%	77.6%	1.89
MO1-2_2826	93,709	-	-	Plasmid	Plasmid	COG3505, COG3451, COG3451, COG3843, COG2161, COG1475, COG1475	Novel	-	-	-	-	-
MO1-2_2827	31,029	-	-	Plasmid	Plasmid	COG5527, COG3505, COG3451		NC_004703.1	Bacteroides thetaiotaomicron VPI-5482 plasmid p5482	99.0%	91.7%	0.94
MO1-2_2829	11,301	-	-	Unclassified	Plasmid	0	Novel	-	-	-	-	-
MO1-2_2831	37,533	-	-	Plasmid	Plasmid	COG1475	Novel	-	-	-	-	-
YS1-2_000081F	208,826	-	-	Chromosome	Plasmid	COG3451, COG1475, COG3505	Novel	-	-	-	-	-
YS1-2_000084F	185,698	-	-	Chromosome	Plasmid	COG3077, COG2161	Novel	-	-	-	-	-
YS1-2_000086F	179,466	-	-	Plasmid	Plasmid	COG3505, COG1475, COG2026, COG2161, COG1475, COG1475, COG3451		NZ_CP008842.1	Klebsiella oxytoca strain M1 plasmid pKOXM1A	99.1%	77.2%	0.88
YS1-2_2427	7,669	-	-	Plasmid	Plasmid	COG5527	Novel	-	-	-	-	-
YS1-2_2428	129,326	-	-	Plasmid	Plasmid	COG1475, COG3451, COG3505, COG1475	Novel	-	-	-	-	-
YS1-2_2430	10,473	-	-	Unclassified	Plasmid	COG3077	Novel	-	-	-	-	-

Name	Length in bp	Classification				Plasmid-enriched genes**	Novelity	High-similar reference ID	High-similar references in databases	Identity with closest reference	Alignment coverage	Contig length / Reference length
		POG*	Virsorter*	PlasFlow	This study							
YS1-2_2431	31,060	-	-	Plasmid	Plasmid	COG3451, COG3505, COG5527		NC_004703.1	Bacteroides thetaiotaomicron VPI-5482 plasmid p5482	98.8%	91.7%	0.94
YS1-2_2432	43,779	-	-	Chromosome	Plasmid	COG3077	Novel	-	-	-	-	-
YS1-2_2434	30,246	1	2	Chromosome	Phage	-	Novel	-	-	-	-	-
YS1-2_2435	104,122	-	-	Plasmid	Plasmid	COG5527, COG1475, COG1475, COG3451, COG3505	Novel	-	-	-	-	-
YS1-2_2437	98,907	1	2	Chromosome	Phage	-		NC_024711.1	Uncultured phage crAssphage	97.3%	94.4%	1.02
									phage:3300009687_Ga0116144_10011271	98.1%	7.0%	14.23
									phage: 3300012956_Ga0154020_10017436	98.3%	8.1%	12.30
									phage: 3300009692_Ga0116171_10015780	98.9%	5.8%	17.15
									phage: 7000000075_SRS016056_WUGC_scaffold_1507	96.8%	8.6%	9.66
									phage: 3300008360_Ga0114875_1003073	98.0%	11.3%	8.87
									phage: 3300007312_Ga0104941_101252	97.2%	9.9%	7.22

Table S2.5: Functional annotation of the 71 plasmid CCs based on COGs. (Table S2.5 is omitted from this thesis due to the large size. It is available at https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0737-z.)

Table S2.6: Intra-similarity and length of TDRs of crAssphage linear genomes.

crAssphage	TDR (left) in bp	TDR (right) in bp	Mismatch bases	Linear genome size in bp
apr34_000142F	2,171	2,171	1	96,602
ES_ALL_000190F	1,922	1,922	1	99,214
FA1-2_000172F	1,891	1,891	2	98,459
GF1-2_000079F	2,452	2,452	1	99,717
YS1-2_2437	1,938	1,938	0	100,844
NC_024711.1	1,853	1,853	0	98,917

Table S2.7: HQ chromosome bins reconstructed from PacBio contigs. *Bold bins indicate circular contigs. **Identity with the 40 marker genes.

Bin/L-MAG ID*	Subject	Genome Size	N50	# of contigs	Comple-teness	Contami-nation	Reference genome ID	Closest reference genome	Identity**	Putative phylum	5S rRNA#	23S rRNA#	16S rRNA#
apr34_c000003F	apr34	2,335,568	2,335,568	1	100.0	0.8	AAVN00000000	Collinsella aerofaciens ATCC 25986	96.77	Actinobacteria	5	5	5
ES_unbinned_6	ES	2,371,174	2,371,174	1	100.0	2.4	AAVN00000000	Collinsella aerofaciens ATCC 25986	96.65	Actinobacteria	5	5	5
apr38_c000003F	apr38	2,639,692	2,639,692	1	100.0	0.0	ACTL00000000	Eubacterium sp. 3_1_31	81.66	Firmicutes	7	7	7
FAKO02_c000003F	FAKO02	2,218,365	2,218,365	1	99.6	0.2	AEEL00000000	Streptococcus equinus ATCC 700338	99.63	Firmicutes	6	6	6
apr38_c000001F	apr38	2,985,567	2,985,567	1	99.3	2.0	AMEY00000000	Anaerostipes hadrus DSM 3319	99.54	Firmicutes	6	5	5
apr34_c000002F	apr34	2,416,973	2,416,973	1	100.0	0.0	CAWP000000000	Veillonella sp. CAG:933	96.28	Firmicutes	3	3	3
FAKO02_c000001F	FAKO02	2,723,705	2,723,705	1	98.7	0.0	CBIU000000000	Eubacterium sp. CAG:38	99.30	Firmicutes	5	5	5
FAKO27_unbinned_1	FAKO27	2,034,231	2,034,231	1	98.2	0.2	JEOD00000000	Bifidobacterium pseudocatenulatum IPLA36007	99.52	Actinobacteria	5	5	5
apr38_unbinned_3	apr38	2,179,748	2,179,748	1	99.6	0.0	NC_008618	Bifidobacterium adolescentis ATCC 15703	98.68	Actinobacteria	6	6	6
ES_c000001F	ES	2,205,987	2,205,987	1	99.5	0.0	NC_017999	Bifidobacterium bifidum BGN4	99.73	Actinobacteria	5	3	3
apr38_11	apr38	3,025,056	1,569,845	3	99.4	0.0	ACTX00000000	Lachnospiraceae bacterium 9_1_43BFAA	99.79	Firmicutes	6	6	6
FAKO02_35	FAKO02	2,875,235	811,490	4	98.7	0.6	ABWO00000000	Tyzzerella nexilis DSM 1787	99.37	Firmicutes	3	3	3
ES_6	ES	2,815,152	974,829	4	99.3	0.7	AECU00000000	Faecalibacterium cf. prausnitzii KLE1255	99.18	Firmicutes	4	4	4
FAKO02_6	FAKO02	2,242,241	1,170,390	4	91.9	0.0	AUYD00000000	Bifidobacterium longum E18	99.65	Actinobacteria	6	7	7
MO1-2_19	MO1-2	2,152,965	875,611	4	96.0	0.3	CAZC000000000	Eubacterium sp. CAG:180	99.36	Firmicutes	5	5	5
FAKO02_47	FAKO02	2,073,102	443,962	4	96.6	2.0	CBEX000000000	Eubacterium sp. CAG:274	97.42	Firmicutes	6	5	5
FAKO02_4	FAKO02	2,302,895	820,342	4	97.3	2.7	JEOD00000000	Bifidobacterium pseudocatenulatum IPLA36007	99.62	Actinobacteria	5	4	4
apr38_20	apr38	3,159,405	951,536	4	100.0	0.0	JMLZ00000000	Robinsoniella sp. KNHs210	79.00	Firmicutes	4	4	4
apr34_15	apr34	2,162,838	1,143,729	4	98.9	0.0	NC_014616	Bifidobacterium bifidum S17	99.59	Actinobacteria	2	2	2
ES_9	ES	3,074,273	833,271	5	98.7	2.4	AMEY00000000	Anaerostipes hadrus DSM 3319	99.50	Firmicutes	5	5	5
ES_17	ES	2,451,186	1,333,983	5	97.6	0.0	CAWP000000000	Veillonella sp. CAG:933	99.23	Firmicutes	5	5	5
apr38_29	apr38	3,410,434	2,571,534	5	99.3	0.0	NC_012778	Eubacterium eligens ATCC 27750	83.71	Firmicutes	5	5	5
apr34_4	apr34	3,429,605	2,116,660	5	99.5	1.0	NC_012781	Eubacterium rectale ATCC 33656	99.37	Firmicutes	7	7	7
FAKO05_6	FAKO05	3,219,832	2,402,855	5	95.2	0.7	NC_021010	Eubacterium rectale DSM 17629	98.93	Firmicutes	4	4	4
FAKO02_21	FAKO02	4,962,458	1,062,471	6	99.4	0.2	AGZN00000000	Parabacteroides distasonis CL09T03C24	99.53	Bacteroidetes	6	6	6
apr34_28	apr34	3,396,077	795,391	6	99.4	0.0	BAHU00000000	Clostridiales bacterium VE202-07	84.78	Firmicutes	8	8	8
FAKO02_25	FAKO02	2,914,187	836,595	6	95.9	0.6	CBIK000000000	Dorea longicatena CAG:42	99.74	Firmicutes	6	6	6
GF1-2_13	GF1-2	2,853,312	654,986	6	99.1	0.0	NC_021020	Faecalibacterium prausnitzii SL3/3	98.53	Firmicutes	6	6	6
YS1-2_15	YS1-2	3,466,694	892,738	6	100.0	0.0	NC_021035	butyrate-producing bacterium SS3/4	98.27	Firmicutes	6	6	6
ES_4	ES	5,577,691	978,158	7	96.2	0.5	AGXJ00000000	Bacteroides dorei CL02T12C06	99.89	Bacteroidetes	5	5	6
apr34_14	apr34	3,082,601	874,590	7	98.3	0.0	AUJS00000000	Dorea longicatena AGR2136	99.16	Firmicutes	7	7	7
GF1-2_23	GF1-2	1,966,580	398,027	7	96.3	1.0	CAZC000000000	Eubacterium sp. CAG:180	99.60	Firmicutes	6	6	6
MO1-2_2	MO1-2	2,173,340	582,374	7	98.6	0.8	JEOD00000000	Bifidobacterium pseudocatenulatum IPLA36007	99.58	Actinobacteria	2	2	1
FAKO02_16	FAKO02	3,265,609	1,769,999	7	99.1	0.1	NC_021015	Ruminococcus torques L2-14	98.58	Firmicutes	6	5	5

Bin/L-MAG ID*	Subject	Genome Size	N50	# of contigs	Comple-teness	Contami-nation	Reference genome ID	Closest reference genome	Identity**	Putative phylum	5S rRNA#	23S rRNA#	16S rRNA#
FAKO02_1	FAKO02	2,263,328	682,757	8	99.6	4.1	AAXD00000000	Bifidobacterium adolescentis L2-32	98.41	Actinobacteria	5	5	6
YS1-2_19	YS1-2	4,652,452	1,041,020	8	95.4	0.0	ABYH00000000	Parabacteroides johnsonii DSM 18315	99.68	Bacteroidetes	5	5	5
FA1-2_21	FA1-2	2,827,177	657,026	8	99.4	0.0	ACTX00000000	Lachnospiraceae bacterium 9_1_43BFAA	99.78	Firmicutes	5	4	4
ES_12	ES	3,104,542	469,338	8	98.0	0.1	NC_021015	Ruminococcus torques L2-14	98.29	Firmicutes	5	4	5
FAKO27_38	FAKO27	2,783,018	1,454,178	9	99.3	0.7	CBBN000000000	Eubacterium sp. CAG:156	99.64	Firmicutes	8	8	8
apr34_32	apr34	2,383,321	1,238,576	9	98.1	0.0	CCMM00000000	bacterium LF-3	95.19	Firmicutes	3	3	3
GF1-2_41	GF1-2	5,332,730	760,826	9	98.9	1.3	JMZA00000000	Klebsiella pneumoniae MGH 65	99.25	Proteobacteria	3	4	3
apr38_1	apr38	2,204,996	540,803	9	95.7	0.0	NC_015067	Bifidobacterium longum subsp. longum JCM 1217	99.22	Actinobacteria	4	4	5
FAKO05_11	FAKO05	2,245,789	299,882	9	98.7	0.7	NC_021013	Ruminococcus bromii L2-63	98.28	Firmicutes	10	8	8
FAKO02_11	FAKO02	2,323,147	592,391	10	99.2	4.0	AAVN00000000	Collinsella aerofaciens ATCC 25986	97.14	Actinobacteria	5	5	3
apr38_4	apr38	3,429,045	923,763	10	93.9	0.6	AAYG00000000	Ruminococcus gnavus ATCC 29149	99.75	Firmicutes	6	6	6
ES_22	ES	3,977,013	521,300	10	96.3	1.1	ABIY00000000	Bacteroides coprocola DSM 17136	99.36	Bacteroidetes	6	5	6
FAKO27_35	FAKO27	2,346,342	327,237	10	98.3	0.6	AEVN00000000	Phascolarctobacterium succinatutens YIT 12067	99.65	Firmicutes	1	1	1
FAKO27_27	FAKO27	2,255,718	560,798	10	94.9	0.3	CAZC000000000	Eubacterium sp. CAG:180	99.27	Firmicutes	4	4	4
GF1-2_38	GF1-2	1,948,170	423,239	10	96.6	0.0	CBFD000000000	Eubacterium sp. CAG:251	99.20	Firmicutes	3	1	2
FAKO27_14	FAKO27	3,417,588	524,621	11	95.6	0.0	AAVO00000000	Ruminococcus obeum ATCC 29174	99.57	Firmicutes	5	5	4
apr34_17	apr34	5,036,842	793,300	11	97.8	0.7	AGZN00000000	Parabacteroides distasonis CL09T03C24	99.08	Bacteroidetes	7	7	7
FAKO02_13	FAKO02	2,310,295	403,800	11	94.9	1.3	ARKD00000000	Megamonas rupellensis DSM 19944	99.52	Firmicutes	9	9	8
ES_31	ES	3,163,092	542,239	11	98.7	0.7	ARTA00000000	Clostridium sporosphaeroides DSM 1294	74.08	Firmicutes	3	3	3
apr34_19	apr34	3,812,964	2,976,383	11	98.7	0.6	BAHU00000000	Clostridiales bacterium VE202-07	83.30	Firmicutes	2	2	1
apr34_1	apr34	3,251,349	500,957	11	91.0	2.2	CBJJ000000000	Blautia sp. CAG:37	96.83	Firmicutes	4	4	4
FAKO02_3	FAKO02	3,676,649	468,092	11	99.0	0.6	CBJJ000000000	Blautia sp. CAG:37	96.93	Firmicutes	5	5	5
apr34_26	apr34	2,209,035	336,180	11	98.2	0.6	NC_016077	Acidaminococcus intestini RyC-MR95	99.99	Firmicutes	6	6	6
ES_5	ES	3,548,018	881,991	12	98.1	0.3	AAYG00000000	Ruminococcus gnavus ATCC 29149	99.81	Firmicutes	8	8	8
FA1-2_8	FA1-2	4,787,664	590,173	12	96.5	2.2	AAYH00000000	Bacteroides uniformis ATCC 8492	99.54	Bacteroidetes	5	5	5
FA1-2_25	FA1-2	3,526,022	385,522	12	93.5	0.0	ABQC00000000	Bacteroides plebeius DSM 17135	98.43	Bacteroidetes	6	6	6
YS1-2_30	YS1-2	6,134,993	1,921,654	12	99.4	0.0	AUUC00000000	Blautia producta ATCC 27340 = DSM 2950	93.35	Firmicutes	5	5	5
apr34_9	apr34	3,494,826	807,729	12	98.7	1.0	AWSY00000000	Blautia sp. KLE 1732	99.39	Firmicutes	4	4	4
FAKO02_8	FAKO02	3,146,426	462,074	12	92.4	1.0	AWSY00000000	Blautia sp. KLE 1732	98.87	Firmicutes	8	8	8
FA1-2_33	FA1-2	3,163,343	378,506	12	92.0	0.1	CBAK000000000	Roseburia sp. CAG:50	98.88	Firmicutes	5	5	6
FAKO27_29	FAKO27	2,850,421	438,689	12	98.7	0.0	NC_012778	Eubacterium eligens ATCC 27750	83.85	Firmicutes	4	4	6
FAKO27_7	FAKO27	3,383,899	354,681	12	98.8	1.5	NC_012781	Eubacterium rectale ATCC 33656	99.24	Firmicutes	5	5	5
YS1-2_21	YS1-2	3,741,847	414,417	13	98.1	0.0	ABFX00000000	Erysipelatoclostridium ramosum DSM 1402	99.93	Firmicutes	7	7	7
GM1-2_18	GM1-2	2,350,187	375,180	13	92.5	4.4	ABYT00000000	Eubacterium biforme DSM 3989	94.06	Firmicutes	7	7	7
GF1-2_7	GF1-2	3,749,671	495,372	13	98.7	4.1	CBCH000000000	Prevotella copri CAG:164	98.75	Bacteroidetes	6	6	6
MO1-2_4	MO1-2	3,739,281	561,749	13	98.6	1.6	CBJJ000000000	Blautia sp. CAG:37	96.22	Firmicutes	5	5	5
MO1-2_7	MO1-2	3,154,245	480,737	13	96.9	1.5	NC_012781	Eubacterium rectale ATCC 33656	99.23	Firmicutes	6	6	6
apr34_7	apr34	5,265,903	493,210	14	97.9	0.4	AGXJ00000000	Bacteroides dorei CL02T12C06	99.91	Bacteroidetes	7	7	7
FAKO02_9	FAKO02	2,715,784	404,903	14	95.0	2.0	AMEY00000000	Anaerostipes hadrus DSM 3319	99.52	Firmicutes	7	6	6

Bin/L-MAG ID*	Subject	Genome Size	N50	# of contigs	Comple-teness	Contami-nation	Reference genome ID	Closest reference genome	Identity**	Putative phylum	5S rRNA#	23S rRNA#	16S rRNA#
ES_35	ES	3,177,248	260,567	14	96.6	0.0	NC_012778	Eubacterium eligens ATCC 27750	83.61	Firmicutes	6	6	6
FAKO27_12	FAKO27	2,671,028	256,091	15	100.0	2.0	ACOP00000000	Faecalibacterium prausnitzii A2-165	96.46	Firmicutes	6	6	6
MO1-2_13	MO1-2	2,112,224	316,718	15	96.7	0.2	NC_017999	Bifidobacterium bifidum BGN4	99.78	Actinobacteria	3	2	2
FA1-2_17	FA1-2	3,257,411	376,941	16	98.3	0.6	AAXB00000000	Dorea longicatena DSM 13814	99.51	Firmicutes	6	6	6
FAKO02_28	FAKO02	2,293,415	207,761	16	92.0	0.0	CAWP000000000	Veillonella sp. CAG:933	99.08	Firmicutes	4	5	4
GM1-2_20	GM1-2	1,890,235	156,125	16	93.3	0.0	CBDU000000000	Clostridium sp. CAG:217	98.49	Firmicutes	3	2	2
FAKO27_19	FAKO27	3,396,119	743,708	17	99.2	2.7	ABYJ00000000	Roseburia intestinalis L1-82	80.34	Firmicutes	4	4	4
YS1-2_17	YS1-2	6,836,119	854,851	17	98.8	2.0	ACWH00000000	Bacteroides ovatus 3_8_47FAA	99.66	Bacteroidetes	4	5	4
ES_30	ES	2,538,944	205,328	17	92.3	1.3	CBEM000000000	Firmicutes bacterium CAG:41	99.27	Firmicutes	6	5	5
YS1-2_5	YS1-2	4,669,747	363,499	18	98.9	3.5	AAYH00000000	Bacteroides uniformis ATCC 8492	97.40	Bacteroidetes	6	4	6
FAKO03_8	FAKO03	3,294,385	409,397	19	99.0	1.2	NC_012781	Eubacterium rectale ATCC 33656	98.92	Firmicutes	7	6	7
FAKO05_4	FAKO05	5,231,222	562,759	20	97.9	1.9	ABWZ00000000	Bacteroides dorei DSM 17855	99.89	Bacteroidetes	8	7	7
FA1-2_3	FA1-2	4,859,606	550,131	20	96.6	1.0	AGXZ00000000	Bacteroides vulgatus CL09T03C04	99.86	Bacteroidetes	8	7	8
FAKO02_63	FAKO02	2,750,947	246,584	21	94.3	3.7	ABYT00000000	Eubacterium biforme DSM 3989	94.55	Firmicutes	6	6	6
apr34_5	apr34	3,895,308	343,063	21	94.9	0.0	AXVN00000000	Blautia wexlerae DSM 19850	99.10	Firmicutes	8	7	7
FA1-2_28	FA1-2	3,205,392	208,534	21	94.2	0.3	BAHU00000000	Clostridiales bacterium VE202-07	84.66	Firmicutes	5	5	5
FA1-2_1	FA1-2	3,405,420	249,378	21	96.8	1.3	CBJJ000000000	Blautia sp. CAG:37	97.28	Firmicutes	6	6	6
GM1-2_2	GM1-2	3,467,527	220,012	21	94.5	0.6	CBJJ000000000	Blautia sp. CAG:37	97.19	Firmicutes	7	6	7
apr38_7	apr38	4,711,604	307,675	21	95.8	1.2	NC_009615	Parabacteroides distasonis ATCC 8503	99.53	Bacteroidetes	6	5	6
FA1-2_2	FA1-2	4,060,289	397,928	22	98.7	2.5	AXVN00000000	Blautia wexlerae DSM 19850	98.70	Firmicutes	6	6	6
apr38_2	apr38	4,633,010	287,735	24	92.5	3.8	JNHI00000000	Bacteroides vulgatus str. 3775 SL(B) 10 (iv)	99.91	Bacteroidetes	9	8	7
FAKO27_9	FAKO27	2,373,830	152,002	25	98.6	0.2	AECU00000000	Faecalibacterium cf. prausnitzii KLE1255	99.26	Firmicutes	6	6	6
GM1-2_7	GM1-2	2,093,470	189,984	27	90.5	0.0	NC_021013	Ruminococcus bromii L2-63	98.32	Firmicutes	3	2	3
ES_34	ES	3,443,471	251,753	29	92.0	2.5	CAXU000000000	Firmicutes bacterium CAG:24	99.07	Firmicutes	6	6	6
FAKO27_46	FAKO27	3,078,972	158,337	31	94.3	0.9	CBGI000000000	Prevotella sp. CAG:592	99.62	Bacteroidetes	7	4	4
FAKO27_6	FAKO27	5,355,380	216,145	35	97.4	0.6	ASSN00000000	Bacteroides vulgatus dnLKV7	99.68	Bacteroidetes	6	8	6
GM1-2_9	GM1-2	3,014,894	126,768	41	90.6	2.1	NC_021015	Ruminococcus torques L2-14	98.29	Firmicutes	3	3	3
apr34_25	apr34	4,150,314	136,095	45	90.9	4.5	AGZQ00000000	Parabacteroides merdae CL03T12C32	99.70	Bacteroidetes	7	7	7

Table S2.8: Host prediction by similarity search of the 71 plasmid CCs for the public genome database. (Table S2.8 is omitted from this thesis due to the large size. It is available at https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0737-z.)

Table S2.9: Summary of host prediction of the 71 plasmid CCs.

Plasmid-assigned CCs*	Length in bp	Novelity	Host prediction methods				Predcted host
			Similarity search against known plasmids	Similarity search against draft genomes/strains	Co-occurrence in the IGCJ dataset	Methylation motif similarity	
apr34_000060F	243,670	Novel			Blautia, Peptoclostridium, Ruminococcus	Clostridiales bacterium VE202-07	Clostridiales
apr34_000136F	98,299	Novel		Bacteroides, Parabacteroides		Bacteroides dorei CL02T12C06	Bacteroidales
apr34_001180F	5,259	Novel		Tyzzerella, Lachnospiraceae, Clostridiales		Clostridiales bacterium VE202-07	Clostridiales
apr34_1785	4,383	Novel		Tyzzerella, Lachnospiraceae, Clostridiales		Clostridiales bacterium VE202-07	Clostridiales
apr34_1786	5,602	Novel		Clostridium, Clostridiales		Clostridiales bacterium VE202-07	Clostridiales
apr34_1788	165,458	Novel		Bifidobacterium			Bifidobacterium
apr38_000029F	617,950	Novel		Eubacterium		Eubacterium eligens ATCC 27750	Eubacterium
apr38_000077F	216,202	Novel			Blautia, Peptoclostridium, Ruminococcus		Clostridiales
apr38_2079	28,607	Novel					
apr38_2081	47,897	Novel					
ES_ALL_000351F	44,025		Bacteroides	Bacteroides, Parabacteroides		Bacteroides dorei CL02T12C06	Bacteroidales
ES_ALL_5057	31,059		Bacteroides	Bacteroides, Parabacteroides			Bacteroidales
ES_ALL_5058	65,724	Novel			Blautia		Blautia
ES_ALL_5059	8,278	Novel					
FA1-2_000250F	62,025	Novel					

Plasmid-assigned CCs*	Length in bp	Novelity	Host prediction methods				Predcted host
			Similarity search against known plasmids	Similarity search against draft genomes/strains	Co-occurrence in the IGCJ dataset	Methylation motif similarity	
FA1-2_000589F	26,749	Novel					
FA1-2_2752	5,396	Novel		Bacteroides, Parabacteroides			Bacteroidales
FA1-2_2754	6,124	Novel					
FA1-2_2755	10,436	Novel		Clostridiales			Clostridiales
FA1-2_2756	50,544	Novel					
FA1-2_2758	8,177	Novel				Blautia sp. CAG:37	Blautia
FA1-2_2760	4,306	Novel	Bacteroides	Bacteroides, Parabacteroides			Bacteroidales
FAKO02_000237F	74,873	Novel		Bacteroides			Bacteroides
FAKO02_3061	25,023	Novel				Collinsella aerofaciens ATCC 25986	Collinsella
FAKO02_3062	54,132	Novel				Collinsella aerofaciens ATCC 25986	Collinsella
FAKO02_3063	70,284	Novel					
FAKO03_2022	5,335	Novel		Tyzzerella, Lachnospiraceae, Clostridiales			Clostridiales
FAKO03_2023	2,782		Bacteroides	Bacteroides, Parabacteroides			Bacteroidales
FAKO03_2024	6,599	Novel					
FAKO03_2027	11,603	Novel					
FAKO03_2028	5,022	Novel		Tyzzerella, Lachnospiraceae, Clostridiales			Clostridiales
FAKO03_2030	3,680		Bifidobacterium	Bifidobacterium			Bifidobacterium
FAKO05_000706F	6,322	Novel				Ruminococcus sp. CAG:330	Ruminococcus

Plasmid-assigned CCs*	Length in bp	Novelity	Host prediction methods				Predcted host
			Similarity search against known plasmids	Similarity search against draft genomes/strains	Co-occurrence in the IGCJ dataset	Methylation motif similarity	
FAKO05_2266	4,970	Novel		Bacteroides		Bacteroides dorei DSM 17855	Bacteroides
FAKO05_2267	4,787	Novel		Bacteroides		Bacteroides dorei DSM 17855	Bacteroides
FAKO05_2268	4,148	Novel	Bacteroides	Bacteroides		Bacteroides dorei DSM 17855, Ruminococcus sp. CAG:330	Bacteroides
FAKO05_2270	3,842	Novel				Blautia wexlerae DSM 19850	Blautia
FAKO05_2271	4,303	Novel	Bacteroides	Bacteroides, Parabacteroides			Bacteroidales
FAKO05_2273	5,594		Bacteroides	Bacteroides			Bacteroides
FAKO05_2274	2,784		Bacteroides	Bacteroides, Parabacteroides			Bacteroidales
FAKO27_001080F	12,395	Novel			Firmicutes bacterium, Oscillibacter (Clostridiales)		Clostridiales
FAKO27_6405	2,873	Novel					
FAKO27_6407	2,970	Novel					
FAKO27_6409	5,023	Novel		Tyzzerella, Lachnospiraceae, Clostridiales			Clostridiales
FAKO27_6410	4,307	Novel	Bacteroides	Bacteroides, Parabacteroides			Bacteroidales
FAKO27_6411	44,683	Novel			Collinsella		Collinsella
FAKO27_6412	7,241	Novel					
FAKO27_6413	2,897	Novel					
GF1-2_000012F	666,740	Novel		Eubacterium	Eubacterium		Eubacterium
GF1-2_000048F	167,388		Lactobacillus	Lactobacillus			Lactobacillus

Plasmid-assigned CCs*	Length in bp	Novelity	Host prediction methods				Predcted host
			Similarity search against known plasmids	Similarity search against draft genomes/strains	Co-occurrence in the IGCJ dataset	Methylation motif similarity	
GF1-2_000127F	72,273	Novel		Clostridium, Erysipelotrichaceae,			Clostridium
GF1-2_000231F	48,134		Clostridium	Clostridium			Clostridium
GF1-2_000286F	37,242	Novel					
GF1-2_000418F	30,976		Bacteroides	Bacteroides, Parabacteroides			Bacteroidales
GF1-2_000511F	26,269	Novel					
GF1-2_4069	117,808	Novel					
GF1-2_4070	7,660	Novel		Bacteroides			Bacteroides
GM1-2_000192F	44,253	Novel	Lactobacillus				Lactobacillus
MO1-2_2826	93,709	Novel					
MO1-2_2827	31,029		Bacteroides	Bacteroides, Parabacteroides			Bacteroidales
MO1-2_2829	11,301	Novel		Bacteroides	Bacteroides		Bacteroides
MO1-2_2831	37,533	Novel					
YS1-2_000081F	208,826	Novel			Blautia, Coprococcus, Peptoclostridium, Ruminococcus	Tyzzerella nexilis DSM 1787	Clostridiales
YS1-2_000084F	185,698	Novel		Bifidobacterium	Bifidobacterium	Bifidobacterium longum subsp. longum ATCC 55813	Bifidobacterium
YS1-2_000086F	179,466		Klebsiella				Klebsiella
YS1-2_2427	7,669	Novel				Tyzzerella nexilis DSM 1787	Clostridiales
YS1-2_2428	129,326	Novel					
YS1-2_2430	10,473	Novel		Clostridiales			Clostridiales
YS1-2_2431	31,060		Bacteroides	Bacteroides, Parabacteroides		Parabacteroides johnsonii DSM 18315	Bacteroidales
YS1-2_2432	43,779	Novel		Clostridium, Clostridiales			Clostridiales
YS1-2_2435	104,122	Novel					

Table S2.10: Summary of host prediction of the 11 phage CCs. *Numbers in parentheses indicate mismach bases with CRISPR spacers.

Phage	Length	High-similar	Host prediction methods				Predicted host
	in bp	known phages	Genomes and contigs with high-similar CRISPRs in the IGCJ dataset*	Genomes with high-similar CRISPRs in the public database*	HQ chromosmal bins with high-similar CRISPRs in the JP PacBio dataset*	Methylation motif similarity in the JP PacBio dataset	
apr34_000142F (linear)	94,435	Uncultured phage crAssphage	Bacteroides dorei 2G11 (0)	Porphyromonas sp. 31_2 (0)			Bacteroidales
			1 unknown (1)				
apr34_1784 (circular)	45,225		Bacteroides cellulosilyticus CL02T12C19 (1)	Bacteroides cellulosilyticus CL02T12C19 (0)	Parabacteroides merdae ATCC 43184 (1)	Bacteroides dorei CL02T12C06	Bacteroidales
			Bacteroides dorei CL03T12C01 (1)	Odoribacter splanchnicus DSM 20712 (0)			
			Bacteroides eggerthii 1_2_48FAA (0)				
			Bacteroides massiliensis B84634 = Timone 84634 (0)				
			Bacteroides rodentium JCM 16496 (0)				
			Bacteroides salyersiae WAL 10018 = DSM 18765 = JCM 12988 (1)				
			Bacteroides sp. 14(A) (0)				
			Bacteroides sp. 3_1_33FAA (0 and 1)				
			Bacteroides sp. 4_3_47FAA (1)				
			Bacteroides uniformis CL03T12C37 (0)				
			Bacteroides uniformis str. 3978 T3 ii (1)				
			Odoribacter splanchnicus DSM 20712 (0 and 1)				
			Parabacteroides distasonis str. 3999B T(B) 4 (0)				
			Parabacteroides distasonis str. 3999B T(B) 6 (0 and 1)				
			Parabacteroides merdae ATCC 43184 (0 and 1)				
			Porphyromonas sp. 31_2 (0)				
			Prevotella sp. CAG:279 (0 and 1)				
			13 unknown contig (0 and 1)				
apr34_1792 (circular)	30,579		Bifidobacterium longum subsp. longum 44B (0 and 1)				Bifidobacterium
ES_ALL_000190F (linear)	97,257	Uncultured phage crAssphage	Bacteroides dorei 2G11 (0)	Porphyromonas sp. 31_2 (0)	Bacteroides vulgatus ATCC 8482 (1)		Bacteroidales
			1 unknown contig (1)				
FA1-2_000172F (linear)	96,565	Uncultured phage crAssphage	Bacteroides dorei 2G11 (0)	Porphyromonas sp. 31_2 (0)			Bacteroidales

Phage	Length	High-similar	Host prediction methods				Predicted host
	in bp	known phages	Genomes and contigs with high-similar CRISPRs in the IGCJ dataset*	Genomes with high-similar CRISPRs in the public database*	HQ chromosmal bins with high-similar CRISPRs in the JP PacBio dataset*	Methylation motif similarity in the JP PacBio dataset	
FAKO05_000032F (linear)	83,596		Bacteroides sp. 3_1_33FAA (0 and 1)	Bacteroides vulgatus CAG:6 (1)			Bacteroides
			Bacteroides sp. 9_1_42FAA (1)				
			Bacteroides vulgatus str. 3975 RP4 (0 and 1)				
			1 unknown contig (0)				
FAKO27_000238F (Circular)	64,223		Phascolarctobacterium succinatutens CAG:287 (0)			Phascolarctobacterium succinatutens YIT 12067	Phascolarctobacterium
			Phascolarctobacterium succinatutens YIT 12067 (1)				
			1 unknown contig (1)				
FAKO27_000271F (linear)	56,426		Faecalibacterium cf. prausnitzii KLE1255 (0 and 1)	Faecalibacterium sp. CAG:82 (0)	Faecalibacterium prausnitzii L2-6 (1)	Faecalibacterium cf. prausnitzii KLE1255	Faecalibacterium
			Faecalibacterium prausnitzii A2-165 (1)				
			Faecalibacterium prausnitzii L2-6 (0 and 1)				
			21 unknown contigs (0 and 1)				
GF1-2_000079F (linear)	97,820	Uncultured phage crAssphage	Bacteroides dorei 2G11 (0)	Porphyromonas sp. 31_2 (0)	Bacteroides vulgatus ATCC 8482 (0)		Bacteroidales
			1 unknown contig (1)				
YS1-2_2434 (circular)	30,246		Bifidobacterium longum CAG:69 (0)	Bifidobacterium longum subsp. longum 44B (0)	Bifidobacterium longum subsp. infantis CCUG 52486 (0)	Bifidobacterium longum subsp. longum ATCC 55813	Bifidobacterium
			Bifidobacterium longum subsp. longum KACC 91563 (0)		Bifidobacterium longum subsp. longum F8 (1)		
			1 unknown contig (1)				
YS1-2_2437 (linear)	98,907	Uncultured phage crAssphage	Bacteroides dorei 2G11 (0)	Porphyromonas sp. 31_2 (0)			Bacteroidales
			1 unknown contig (1)				

Table S2.11: Clusters of plasmids and phages, putative hosts, and the number of reads mapped to the clusters in the IGCJ dataset. (Table S2.11 is omitted from this thesis due to the large size. It is available at `https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0737-z`.)

Table S2.12: Estimation of ratio of plasmids and crAssphage per microbial chromosome in the IGCJ dataset. *The ratio of reads mostly mapped to other phages is not shown in the table.

Country	Ratio of mapped reads*			Average genome size			Estimated ratio per microbial chromosome	
	All plasmid	crAssphage	Chromosome	Plasmid	crAssphage	Chromosome	Plasmid (average)	crAssphage (average)
CN	1.54%	0.23%	98.04%	23,334	99,193	3,870,199	4.27	0.10
DK	0.56%	0.05%	99.15%	26,854	99,193	3,160,834	1.16	0.02
ES	1.01%	0.13%	98.58%	19,125	99,193	3,276,192	2.27	0.04
JP	1.24%	0.03%	98.69%	15,389	99,193	3,178,300	3.23	0.01
US	1.19%	1.40%	96.63%	20,148	99,193	4,115,259	4.03	0.69
All	1.11%	0.38%	98.20%	20,715	99,193	3,516,884	3.03	0.18

Table S2.13: COGs having significant difference in abundance between plasmids and reference genomes detected in the IGCJ dataset. (Table S2.13 is omitted from this thesis due to the large size. It is available at `https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0737-z`.)

Table S2.14: Resfams-based antibiotic resistance functions in plasmids detected in the IGCJ dataset. (Table S2.14 is omitted from this thesis due to the large size. It is available at `https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0737-z`.)

Chapter 3

第3章は5年以内に出版される予定であるため、最長で2025年3月22日までインターネット公表できません。
(Chapter 3 cannot be made public on the Internet until up to March 22nd, 2025, because the relevant content is scheduled to be published within 5 years.)

Chapter 4

Conclusions and discussion

In this thesis, I explored methods and frameworks for sequence assembly of repeats through the two chapters. I presented that long-read sequencing technology was indeed effective for:

1) identification of eMGEs as complete contigs from human gut microbiomes; and

2) characterization of centromeric tandem repeats for better read overlap even within repeats.

Long reads are very useful, but to avoid potential misassemblies and to derive full benefit, methods capable of properly handling sequencing errors are necessary. The situation is same even with recent accurate circular consensus reads if we wish to reconstruct highly repetitive regions, as we saw in Chapter 3.

In Chapter 2, using noisy long reads, I could not obtain strain-level contigs. That is, the assembled contigs might be consensus sequences of multiple distinct strains, and some diverged strains were fragmented. Strategies employed in recent phasing algorithms, especially those for highly divergent regions (Kajitani et al., 2019), should be applied to this problem because the difference between strains is often such a situation. Another promising approach is to use circular consensus reads, and some researchers have already reported it.

I developed a distinct method for each of eMGEs (interspersed repeats) and complex satellites (tandem repeats) so far. The conditions for perfect assembly I introduced in Chapter 1 are quite simple, but no one has achieved an ideal, integrated genome assembly including automated repeat assembly. To my best knowledge, even the objective function (i.e., accurate formulation) for the ultimate genome assembly is largely unknown yet.

The Bayesian approach I proposed in Chapter 3 would be one step for perfect assembly. To be honest, however, I believe more improvements are needed for better assembly other than those I described in the Conclusions and Discussion section in the chapter. Although I indeed overcame the limitation of the frequentist approach, i.e., single-vs-single read overlap, by introducing a background genome model, I still rely on pairwise overlaps in the end. There is actually an inevitable limit of the approach due to sequencing errors, i.e., some of the pairwise read overlaps could fail even if I employ an appropriate genome model. Removing such small portion of false overlaps in the layout step is one workaround, but I doubt if it would be non-optimal heuristics. Therefore, I believe direct inference of the genome model from reads without the overlap step is required ultimately. Another issue is the genome model. In Chapter 3, I assumed the background sequence as a set of unit sequences and developed a mixture model specialized for tandem repeats. However, genome sequences are in fact a set of single strings, and the model should also be ideally so. In addition, I believe that information on the read depth of the paths in the assembly graph is essential. This fact might be very trivial at least for genome assembly researchers, but I believe I need to exploit the depth information more.

Bibliography

Abouelhoda, M. I. and Ohlebusch, E. (2005). Chaining algorithms for multiple genome comparison. *Journal of Discrete Algorithms*, 3(2):321 – 341. 12

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410. 13

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402. 12

Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., and Pevzner, P. A. (2016). plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*, 32(22):3380–3387. 23, 35

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180. 9, 22

Beaulaurier, J., Zhang, X.-S., Zhu, S., Sebra, R., Rosenbluh, C., Deikus, G., Shen, N., Munera, D., Waldor, M. K., Chess, A., et al. (2015). Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nature Communications*, 6(1):7438. 11

Beaulaurier, J., Zhu, S., Deikus, G., Mogno, I., Zhang, X.-S., Davis-Richardson, A., Canepa, R., Triplett, E. W., Faith, J. J., Sebra, R., et al. (2018). Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nature Biotechnology*, 36(1):61–69. 19, 26, 35

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., and Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6):623–630. 13

Bishara, A., Moss, E. L., Kolmogorov, M., Parada, A. E., Weng, Z., Sidow, A., Dekas, A. E., Batzoglou, S., and Bhatt, A. S. (2018). High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nature Biotechnology*, 36:1067–1075. 15, 19

Bongartz, P. (2019). Resolving repeat families with long reads. *BMC Bioinformatics*, 20(1):232. 15

Bongartz, P. and Schloissnig, S. (2018). Deep repeat resolution—the assembly of the Drosophila histone complex. *Nucleic Acids Research*, 47(3):e18. 15

Bresler, G., Bresler, M., and Tse, D. (2013). Optimal assembly for high throughput shotgun sequencing. *BMC Bioinformatics*, 14(5):S18. 14, 15

Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., Naisilisili, W., Tamminen, M., Smillie, C. S., Wortman, J. R., et al. (2016). Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 535(7612):435–439. 19, 45

Brown, B. L., Watson, M., Minot, S. S., Rivera, M. C., and Franklin, R. B. (2017). MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. *GigaScience*, 6(3). 19

Castro-Mejía, J., Muhammed, M. K., Kot, W., Neve, H., Franz, C. M. A. P., Hansen, L. H., Vogensen, F. K., and Nielsen, D. S. (2015). Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome*, 3(64). 19

Chaisson, M. J. P., Wilson, R. K., and Eichler, E. E. (2015). Genetic variation and the *de novo* assembly of human genomes. *Nature Reviews Genetics*, 16(11):627–640. 10

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13(12):1050–1054. 10, 15, 21, 42

Chung, C.-H., Walter, M. H., Yang, L., Chen, S.-C. G., Winston, V., and Thomas, M. A. (2017). Predicting genome terminus sequences of *Bacillus* cereus-group bacteriophage using next generation sequencing data. *BMC Genomics*, 18(1):350. 32

Churchill, G. A. and Waterman, M. S. (1992). The accuracy of DNA sequences: Estimating sequence quality. *Genomics*, 14(1):89–98. 8

Cornuault, J. K., Petit, M.-A., Mariadassou, M., Benevides, L., Moncaut, E., Langella, P., Sokol, H., and De Paepe, M. (2018). Phages infecting *Faecalibacterium prausnitzii* belong to novel viral genera that help to decipher intestinal viromes. *Microbiome*, 6(1):65. 44

Coyne, M. J., Zitomersky, N. L., McGuire, A. M., Earl, A. M., and Comstock, L. E. (2014). Evidence of extensive DNA transfer between *Bacteroidales* species within the human gut. *mBio*, 5(3):e01305–14. 45

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, pages 345–352. 12

Dib, J. R., Wagenknecht, M., Farías, M. E., and Meinhardt, F. (2015). Strategies and approaches in plasmidome studies—uncovering plasmid diversity disregarding of linear elements? *Frontiers in Microbiology*, 6(463). 19, 44

Duranti, S., Lugli, G. A., Mancabelli, L., Armanini, F., Turroni, F., James, K., Ferretti, P., Gorfer, V., Ferrario, C., Milani, C., et al. (2017). Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. *Microbiome*, 5(1):66. 44

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. 12

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., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications*, 5(1):4498. 24, 31, 35, 43

Edgar, R. C. (2007). PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, 8(1):18. 26

Edwards, R. A., McNair, K., Faust, K., Raes, J., and Dutilh, B. E. (2015). Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiology Reviews*, 40(2):258–272. 35, 44

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138. 10

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl_2):W29–W37. 27

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., and Turner, S. W. (2010). Direct detection of dna methylation during single-molecule, real-time sequencing. *Nature Methods*, 7(6):461–465. 10

Frank, J. A., Pan, Y., Tooming-Klunderud, A., Eijsink, V. G. H., McHardy, A. C., Nederbragt, A. J., and Pope, P. B. (2016). Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Scientific Reports*, 6(25373). 19

Frith, M. C. (2019). How sequence alignment scores correspond to probability models. *Bioinformatics*. 12

Frith, M. C., Wan, R., and Horton, P. (2010). Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Research*, 38(7):e100. 12

Gibson, M. K., Forsberg, K. J., and Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME Journal*, 9(1):207–216. 27, 40

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–1359. 18

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):705–708. 11

Guerin, E., Shkoporov, A., Stockdale, S. R., Clooney, A. G., Ryan, F. J., Sutton, T. D. S., Draper, L. A., Gonzalez-Tortuero, E., Ross, R. P., and Hill, C. (2018). Biology and taxonomy of crAss-like Bacteriophages, the most abundant virus in the human gut. *Cell Host & Microbe*, 24(5):653–664.e6. 43

Guy, L., Roat Kultima, J., and Andersson, S. G. E. (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics*, 26(18):2334–2335. 24

Hamada, M., Wijaya, E., Frith, M. C., and Asai, K. (2011). Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection. *Bioinformatics*, 27(22):3085–3092. 12

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919. 12

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119. 22, 27

Idury, R. M. and Waterman, M. S. (1995). A new algorithm for DNA sequence assembly. *Journal of Computational Biology*, 2(2):291–306. 9

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. 9

Ishiura, H., Shibata, S., Yoshimura, J., Suzuki, Y., Qu, W., Doi, K., Almansour, M. A., Kikuchi, J. K., Taira, M., Mitsui, J., et al. (2019). Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nature Genetics*, 51(8):1222–1232. 14

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, 12(4):351–356. 10, 13

Jain, M., Olsen, H. E., Turner, D. J., Stoddart, D., Bulazel, K. V., Paten, B., Haussler, D., Willard, H. F., Akeson, M., and Miga, K. H. (2018). Linear assembly of a human centromere on the y chromosome. *Nature Biotechnology*, 36(4):321–323. 15

Jørgensen, T. S., Kiil, A. S., Hansen, M. A., Sørensen, S. J., and Hansen, L. H. (2015). Current strategies for mobilome research. *Frontiers in Microbiology*, 5(750). 19

Kajitani, R., Yoshimura, D., Okuno, M., Minakuchi, Y., Kagoshima, H., Fujiyama, A., Kubokawa, K., Kohara, Y., Toyoda, A., and Itoh, T. (2019). Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nature Communications*, 10(1):1702. 15, 83

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165. 21, 23

Kasahara, M. and Morishita, S. (2006). *Large-Scale Genome Sequence Processing*. Imperial College Press. 12

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780. 23

Kececioglu, J. D. and Myers, E. W. (1995). Combinatorial algorithms for DNA sequence assembly. *Algorithmica*, 13(1):7–51. 8

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493. 13

Kim, M.-S., Park, E.-J., Roh, S. W., and Bae, J.-W. (2011). Diversity and abundance of single-stranded DNA viruses in human feces. *Applied and Environmental Microbiology*, 77(22):8062–8070. 44

Kim, S.-W., Suda, W., Kim, S., Oshima, K., Fukuda, S., Ohno, H., Morita, H., and Hattori, M. (2013). Robustness of gut microbiota of healthy adults in response to probiotic intervention revealed by high-throughput pyrosequencing. *DNA Research*, 20(3):241–253. 20

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5):540–546. 10

Koonin, E. V. and Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, 36(21):6688–6719. 19

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research*, 27(5):722–736. 10, 13

Krawczyk, P. S., Lipinski, L., and Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, 46(6):e35–e35. 24, 31

Kristensen, D. M., Cai, X., and Mushegian, A. (2011). Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *Journal of Bacteriology*, 193(8):1806–1814. 23, 31

Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., and Snyder, M. (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nature Biotechnology*, 34(1):64–69. 19

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V. K., Srivastava, T. P., et al. (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Research*, 14(4):169–181. 18

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):R12. 24, 25

Lander, E. S. and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3):231–239. 8

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359. 21

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541–546. 25

Letunic, I. and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, 44(W1):W242–W245. 23, 24

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. 12

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676. 22, 26

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100. 13

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J. R., Prifti, E., Nielsen, T., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, 32(8):834–841. 19, 21, 25, 35

Manrique, P., Bolduc, B., Walk, S. T., van der Oost, J., de Vos, W. M., and Young, M. J. (2016). Healthy human gut phageome. *Proceedings of the National Academy of Sciences*, 113(37):10400–10405. 19

Mende, D. R., Sunagawa, S., Zeller, G., and Bork, P. (2013). Accurate and universal delineation of prokaryotic species. *Nature Methods*, 10(9):881–884. 35

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., and Antin, P. (2016). The iPlant collaborative: Cyberinfrastructure for enabling data to discovery for the life sciences. *PLOS Biology*, 14(1):e1002342. 24

Miller, W. and Myers, E. W. (1988). Sequence comparison with concave weighting functions. *Bulletin of Mathematical Biology*, 50(2):97–120. 11

Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2013). Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences*, 110(30):12450–12455. 19

Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2012). Hypervariable loci in the human gut virome. *Proceedings of the National Academy of Sciences*, 109(10):3962–3966. 19

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2011). The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Research*, 21(10):1616–1625. 19

Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E., and Ghai, R. (2013). Expanding the marine virosphere using metagenomics. *PLOS Genetics*, 9(12):e1003987. 24, 26

Myers, E. W. (1986). An *O(ND)* difference algorithm and its variations. *Algorithmica*, 1(1):251–266. 12

Myers, E. W. (1995). Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, 2(2):275–290. 8, 9

Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, 21(suppl_2):ii79–ii85. 8, 9

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H. J., Remington, K. A., et al. (2000). A whole-genome assembly of *Drosophila*. *Science*, 287(5461):2196–2204. 8, 9

Myers, G. (1999). A fast bit-vector algorithm for approximate string matching based on dynamic programming. *Journal of the ACM*, 46(3):395–415. 12

Myers, G. (2013). *What's Behind Blast*, pages 3–15. Springer London, London. 13

Myers, G. (2014). Efficient local alignment discovery amongst noisy long reads. In Brown, D. and Morgenstern, B., editors, *Algorithms in Bioinformatics*, pages 52–67, Berlin, Heidelberg. Springer Berlin Heidelberg. 12, 13, 21

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Research*, 40(20):e155. 9, 15

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453. 11

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D. R., Gautier, L., Pedersen, A. G., Le Chatelier, E., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8):822–828. 19

Nishijima, S., Suda, W., Oshima, K., Kim, S. W., Hirose, Y., Morita, H., and Hattori, M. (2016). The gut microbiome of healthy japanese and its microbial and functional uniqueness. *DNA Research*, 23(2):125–133. 19, 20, 21, 22, 23, 25, 27, 28, 35

Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A. W. C., Pippel, M., Winkler, S., Hastie, A. R., Young, G., Roscito, J. G., et al. (2018). The axolotl genome and the evolution of key tissue formation regulators. *Nature*, 554(7690):50–55. 10

Ogilvie, L. A., Firouzmand, S., and Jones, B. V. (2012). Evolutionary, ecological and biotechnological perspectives on plasmids resident in the human gut mobile metagenome. *Bioengineered*, 3(1):13–31. 45

Paez-Espino, D., Chen, I.-M. A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V. M., Nielsen, T., et al. (2016). IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Research*, 45(D1):D457–D465. 24, 31

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693. 24

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055. 23

Patel, R., Lin, M., Laney, M., Kurn, N., Rose, S., and Ullman, E. F. (1996). Formation of chimeric DNA primer extension products by template switching onto an annealed downstream oligonucleotide. *Proceedings of the National Academy of Sciences*, 93(7):2969–2974. 43

Patterson, M., Marschall, T., Pisanti, N., van Iersel, L., Stougie, L., Klau, G. W., and Schönhuth, A. (2015). WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology*, 22(6):498–509. 15

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448. 13, 23

Peltola, H., Söderlund, H., and Ukkonen, E. (1984). SEQAID: a DNA sequence assembling program based on a mathematical model. *Nucleic Acids Research*, 12(1):307–321. 8

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753. 9

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65. 18, 19, 25

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60. 25

Reyes, A., Haynes, M., Hanson, N., Angly, F. E., Heath, A. C., Rohwer, F., and Gordon, J. I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, 466(7304):334–338. 19

Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F., and Gordon, J. I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. *Nature Reviews Microbiology*, 10(9):607–617. 19, 44

Roach, J. C., Boysen, C., Wang, K., and Hood, L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, 26(2):345–353. 9

Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015a). VirSorter: mining viral signal from microbial genomic data. *PeerJ*, 3:e985. 23, 24, 31

Roux, S., Hallam, S. J., Woyke, T., and Sullivan, M. B. (2015b). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*, 4:e08490. 31, 34

San Millan, A. (2018). Evolution of plasmid-mediated antibiotic resistance in the clinical context. *Trends in Microbiology*, 26(12):978–985. 45

Sanger, F., Coulson, A., Hong, G., Hill, D., and Petersen, G. (1982). Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology*, 162(4):729–773. 8

Sankoff, D. (1972). Matching sequences under deletion/insertion constraints. *Proceedings of the National Academy of Sciences*, 69(1):4–6. 11

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., and Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6):461–468. 12

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069. 24

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM Journal on Applied Mathematics*, 26(4):787–793. 11, 12

Sharon, I., Kertesz, M., Hug, L. A., Pushkarev, D., Blauwkamp, T. A., Castelle, C. J., Amirebrahimi, M., Thomas, B. C., Burstein, D., Tringe, S. G., et al. (2015). Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Research*, 25(4):534–543. 19

Shkoporov, A. N., Ryan, F. J., Draper, L. A., Forde, A., Stockdale, S. R., Daly, K. M., McDonnell, S. A., Nolan, J. A., Sutton, T. D. S., Dalmasso, M., et al. (2018). Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome*, 6(68). 19

Silver, S. and Walderhaug, M. (1992). Gene regulation of plasmid- and chromosome-determined inorganic ion transport in bacteria. *Microbiology and Molecular Biology Reviews*, 56(1):195–228. 45

Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C., and de la Cruz, F. (2010). Mobility of plasmids. *Microbiology and Molecular Biology Reviews*, 74(3):434–452. 45

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197. 11

Šošić, M. and Šikić, M. (2017). Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395. 12

Stern, A., Mick, E., Tirosh, I., Sagy, O., and Sorek, R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Research*, 22(10):1985–1994. 35

Stewart, R. D., Auffret, M. D., Warr, A., Wiser, A. H., Press, M. O., Langford, K. W., Liachko, I., Snelling, T. J., Dewhurst, R. J., Walker, A. W., et al. (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Communications*, 9(1):870. 44

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445. 27

Suzuki, Y., Korlach, J., Turner, S. W., Tsukahara, T., Taniguchi, J., Qu, W., Ichikawa, K., Yoshimura, J., Yurino, H., Takahashi, Y., et al. (2016). AgIn: measuring the landscape of CpG methylation of individual repetitive elements. *Bioinformatics*, 32(19):2911–2919. 10

Suzuki, Y., Nishijima, S., Furuta, Y., Yoshimura, J., Suda, W., Oshima, K., Hattori, M., and Morishita, S. (2019). Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. *Microbiome*, 7(119). 3

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, 30(12):2725–2729. 23

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214. 18, 19, 25

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*, 282(5396):2012–2018. 9

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2012). Integrative Genomics

Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192. 24

Tischler-Höhle, G. (2019). Haplotype and repeat separation in long reads. In Bartoletti, M., Barla, A., Bracciali, A., Klau, G. W., Peterson, L., Policriti, A., and Tagliaferri, R., editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 103–114, Cham. Springer International Publishing. 14, 15

Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., and Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, 38(15):e159. 11

Tsai, Y. C., Conlan, S., Deming, C., Segre, J. A., Kong, H. H., Korlach, J., and Oh, J. (2016). Resolving the complexity of human skin metagenomes using single-molecule sequencing. *mBio*, 7(1):e01948–15. 19

Turner, J. S. (1989). Approximation algorithms for the shortest common superstring problem. *Information and Computation*, 83(1):1–20. 8

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., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43. 9

Ueno, M., Kikuchi, M., Oshima, K., Kim, S., Morita, H., and Hattori, M. (2011). *Assessment and Improvement of Methods for Microbial DNA Preparation from Fecal Samples*, pages 191–198. John Wiley and Sons. 20, 21

Ukkonen, E. (1985). Algorithms for approximate string matching. *Information and Control*, 64(1):100–118. 12

Ukkonen, E. (1990). A linear-time algorithm for finding approximate shortest common superstrings. *Algorithmica*, 5(1):313–323. 8

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351. 9

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74. 9

Virgin, H. W. (2014). The virome in mammalian physiology and disease. *Cell*, 157(1):142–150. 19

Vollger, M. R., Dishuck, P. C., Sorensen, M., Welch, A. E., Dang, V., Dougherty,

M. L., Graves-Lindsay, T. A., Wilson, R. K., Chaisson, M. J. P., and Eichler, E. E. (2019). Long-read sequence and assembly of segmental duplications. *Nature Methods*, 16(1):88–94. 15

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE*, 9(11):e112963. 22

Waterman, M. S. (1984). Efficient sequence alignment algorithms. *Journal of Theoretical Biology*, 108(3):333–337. 11

Weber, J. L. and Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome Research*, 7(5):401–409. 9

Weirather, J., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X., Buck, D., and Au, K. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, 6(100). 10

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162. 11

Wu, S., Manber, U., Myers, G., and Miller, W. (1990). An *O(NP)* sequence comparison algorithm. *Information Processing Letters*, 35(6):317–323. 12

Yoshimura, J., Ichikawa, K., Shoura, M. J., Artiles, K. L., Gabdank, I., Wahba, L., Smith, C. L., Edgley, M. L., Rougvie, A. E., Fire, A. Z., et al. (2019). Recompleting the *Caenorhabditis elegans* genome. *Genome Research*, 29(6):1009–1022. 15

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Research*, 38(12):e132. 24