RecurM: a Novel Approach for Plasmid Discovery in Metagenomics

Daniel Rawlinson (42046660)

Australian Centre for Ecogenomics

A Research Report submitted for the degree of Bachelor of Science (Honours) at the University of Queensland in October 2019

Molecular Biosciences Honours (BIOC6511)

Primary Supervisor: Gene Tyson

Co-Supervisor: Ben Woodcroft

Word Count: 7758

Declaration by author

This research report is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text.

I have clearly stated the contribution by others to jointly-authored works that I have included in my report. I have clearly stated the contribution of others to my research report as a whole including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my report. The content of my report is the result of work I have carried out since the commencement of my honours research project and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my research report, if any, have been submitted to qualify for another award.

I acknowledge that copyright of all material contained in my research report resides with the copyright holder(s) of that material.

Statement of Contributions to Jointly Authored Works Contained in the Research Report

No jointly-authored works.

Statement of Contributions by Others to the Research Report as a Whole

Metagenomes used in this project were assembled by Dylan Cronin at Ohio State University.

Statement of Parts of the Research Report or Submitted to Qualify for the Award of Another Degree

None.

Published Works by the Author Incorporated into the Research

None.

Additional Published Works by the Author Relevant to the Research Report but not Forming Part of it

None.

Acknowledgements

I wish to express my sincere appreciation to Gene Tyson and Ben Woodcroft for their patience and wisdom while supervising this project. The opportunity to learn from them over the course of this year has spurred my scientific interest immensely.

I am also grateful to all at ACE for providing such a comfortable and welcoming environment in which to practice science. I owe this project's achievements to all of you.

Signature of Author: __________________________ **Date:** 21/10/19

Principal Supervisor Agreement

I have read the final report and agree with the student's declaration.

Signature of Principal Supervisor: G. Tyson **Date:** 21/10/19

Summary

As sequencing technologies mature, more information than ever is becoming available to biology researchers. In metagenomics, the enormous amount of genetic data sequenced from entire microbial communities can be parsed to reveal new information about the composition and function of microbial communities. However, metagenomics currently focuses on core microbial genomes and is of limited power in assessing the pool of Mobile Genetic Elements (MGEs) in a microbiome that contribute significantly to overall microbiome function.

Plasmids, as a key component of MGEs, are routinely neglected in bioinformatic microbiome analysis owing to a deficit in effective tools for their recovery. The difficulty in finding plasmids is especially acute because of their enormous heterogeneity and plasticity. Plasmid databases are biased towards clinically relevant and culturable host microorganisms, which further constrains the variety of plasmids that can be recovered from environments. The development of more sophisticated tools for plasmid discovery is essential in advancing our ability to explain how environments are shaped and impacted by microbes.

This project pioneers a new method for the discovery of plasmids from metagenomic sequence data. Recurrent Assembly, here encoded in a tool called RecurM, searches for identical or near-identical contigs across numerous metagenomes. It is presently shown that sequences repeatedly materialising in the same format are enriched for plasmid contents and may constitute complete plasmid units. A set of 377 assembled metagenomes from thawing permafrost in Northern Sweden were analysed with this method to retrieve 4,267 putative plasmids. Furthermore, the substantial collection of metagenomic data input into RecurM can be leveraged to make inferences as to the possible hosts of these putative plasmids. Correlated relative abundance with 630 high quality MAGs allowed 466 of these putative plasmids to be associated with their microbial host(s).

With development of laboratory methods alongside RecurM for confirmation of results, the bioinformatic approach detailed here stands to make a significant contribution to our understanding of the diversity and functions of plasmids in microbial ecology.

Contents

Abbreviations & Key Terms

RAEs – Repeatedly Assembled Elements

ACE – Australian Centre for Ecogenomics

HGT – Horizontal Gene Transfer

bp – base pair

MAG – Metagenome Assembled Genome

PP – Putative Plasmid

MGEs - Mobile Genetic Elements

Nucmer – A nucleotide-based genome alignment tool packaged in the Mummer suite of software for genome alignments

Cluster – A collection of identical or near-identical contigs

Representative Sequence – The sequence in a cluster with the greatest number of direct alignments
Length – The nucleotide length of the Representative Sequence
Magnitude – The number of sequences in a cluster

Threshold Metrics – A set of three calculations used to determine repeated assembly

LR – Length Ratio
RA – Ratio of Alignment
ANI – Average Nucleotide Identity

Section 1 - Introduction

1.1 Plasmid Introduction

Plasmids are extrachromosomal genetic elements that are stably maintained alongside the chromosome of a microbial host. They are horizontally transferred between and within species, and transferred vertically to a microorganism's offspring as part of its collective genome (Lorenzo-Díaz *et al.*, 2017). They are widespread in bacteria and archaea but have also been discovered in unicellular eukaryotes (Gunge *et al.*, 1982, Blaisonneau *et al.*, 1999).

The genetic content of plasmids is varied, but they are frequently studied for the dissemination of antimicrobial resistance (AMR) through populations and environments (Tyagi *et al.*, 2019, Jitwasinkul *et al.*, 2016, de Been *et al.*, 2014, Gupta *et al.*, 2018). Other functions that can be acquired through plasmid-mediated Horizontal Gene Transfer (HGT) include virulence traits (Johnson & Nolan, 2009, Bohm *et al.*, 2015), heavy metal resistance (Baker-Austin *et al.*, 2006) and genes that enhance the strength of symbiotic relationships (Brom *et al.*, 1992, Pistorio *et al.*, 2008). The breadth of genetic material carried and transferred by plasmids makes them a key vehicle of genetic innovation and host evolution (Sentchilo *et al.*, 2013).Understanding plasmid ecology is thus an essential step in understanding microbial ecology overall.

1.2 Plasmid Biology & Typing

The study of plasmids is hampered by their heterogeneity and the extraordinary capacity they possess to move genetic material within and between themselves. For instance, most known plasmids are circular, but linear types also exist which escape detection and characterisation with conventional protocols (see section 1.3; Hayakawa *et al.*, 1979, Hinnebusch & Tilly, 1993). However, not all plasmids are capable of transfer between microorganisms. Conjugative plasmids are self-transmissible and mobilizable plasmids are transmissible only when assisted by another genetic element in the cell. Non-mobilizable plasmids are not capable of co-ordinated transfer at all and can only be mobilized with transduction or natural transformation (Smillie *et al.*, 2010).

Attempts at classifying plasmids into types have exploited several differentiators (Shintani *et al.*, 2015). Incompatibility (Inc) groups are based on the observation that some plasmids cannot be inherited together because of a shared replication system. Replicon (Rep) typing is a more rigorous interpretation of Inc groups based on sequence analysis of the genes encoding replication machinery (Gotz *et al.*, 1996) and is further divisible by plasmid multilocus sequence type (pMLST) (Garcia-Fernandez *et al.*, 2008).

Mobility (MOB) typing is an alternative classification scheme which examines loci encoding mobility functions. Relaxases are essential enzymes for the mobilization of plasmids and are believed to be universal in conjugative and mobilizable types (Orlek *et al.*, 2017a). Alternatively, replicons are not universal but are capable of finer typing resolution (del Solar *et al.*, 1998).

Typing on these schemes remains an inherently difficult task for a variety of reasons. Often, MOB and Rep types complement each other and exhibit conserved pairings, but such pairings are not observed in all cases because of backbone mosaicism - the mixing and matching of mobility and replication loci across plasmid genomes (Orlek *et al.*, 2017a).

There are instances in which plasmids cannot be typed under one or the other scheme. For Rep typing, plasmids from new environments can contain replication regions that are not represented in the replicon typing system (Sobecky *et al.*, 1997). In addition, two replicons can sometimes be found on the one plasmid (Garcillan-Barcia *et al.*, 2009, de Been *et al.*, 2014). For MOB typing, some plasmids are transferrable using machinery that are not covered by the current MOB typing scheme (Smillie *et al.*, 2010), and neither mobility nor replicon typing are equipped to classify plasmids that are non-mobilizable. Furthermore, non-mobilizable plasmids cover the same size range as mobilizable and conjugative plasmids, so determining mobility potential based on size is impossible (Smillie *et al.*, 2010).

Some plasmids encode not just accessory genes but also genes that are essential for the survival of the host. Harrison *et al* (2010) propose the new category of ‘chromid’ for these units, as they exhibit a genomic signature much closer to the chromosome and house some of the core genome (Harrison *et al.*, 2010). Approximately 10% of bacteria are estimated to have their genome split into separate replicons in this manner (Fournes *et al.*, 2018). Some Rhizobia species contain plasmids that together represent 30-50% of the entire genome (Zahran, 2017). The existence of these elements blurs the distinction between extrachromosomal plasmids and the core genome.

Amongst all the variety in plasmid types, genetic material is constantly exchanged between plasmids of different types (Sheppard *et al.*, 2016). Individual genes move across to other plasmids via transposition (Branger *et al.*, 2018), including in plasmids that show no other mechanism of transfer (Fondi *et al.*, 2010). Even individuals within a single strain cannot be expected to share common plasmid genes (Casjens *et al.*, 2002). The genes encoded on a plasmid are thus highly variable and are not conserved across different plasmid types (Yano *et al.*, 2019).

In this context, tracing plasmids responsible for dissemination of a trait is a challenge which may not be achievable without knowing full plasmid DNA sequences (Conlan *et al.*, 2014). Plasmid typing is a placeholder – an exercise of convenience until full plasmid structures can be accessed easily (Garcillan-Barcia *et al.*, 2009).

1.3 Laboratory Methods

Early attempts to characterise plasmids relied on ‘endogenous methods’ requiring cultivation and isolation of a plasmid-carrying microorganism. Endogenous approaches have been used to some success in exploring the basic nature of plasmids (Sobecky *et al.*, 1997, Guerra *et al.*, 2002, Mann *et al.*, 1986), but are limited to the study of plasmids that confer selectable traits and can only be applied to culturable microorganisms.

Later, ‘exogenous methods’ enabled the study of plasmids form non-culturable microorganisms by facilitating the transfer of plasmids to a cultivable recipient (Segura *et al.*, 2014). While the approach has been used extensively to examine the plasmid content of uncultivated environmental samples (Bale *et al.*, 1988, Lilley & Bailey, 1997, Dahlberg *et al.*, 1997), it is incapable of capturing all plasmids in a sample due to its reliance on the plasmid’s mobility potential and compatibility with the surrogate host.

Culture-independent transposon-aided capture (TRACA) increases the output of exogenous plasmids by incorporating an origin of replication and selectable marker into plasmids (Jones & Marchesi, 2007), but is constrained by its capture of mainly small plasmids, and its exclusion of linear plasmids from the process (Dib *et al.*, 2015).

The most high-throughput laboratory method for the characterisation of all plasmids in a sample (the ‘plasmidome’) follows the protocol laid down by Brown Kav *et al.* (2013). Plasmid-safe DNAse is used to break down linear chromosomal DNA segments, followed by treatment with phi29 DNA Polymerase to amplify circular DNA. Enriched plasmid DNA is then directly sequenced, providing a culture-independent metagenomic library of sequence data from an entire population of plasmids.

There are elements of information that whole-plasmidome extraction fails to gather. Primarily, Plasmid-safe DNAse necessarily excludes linear plasmids from the enriched plasmidome (Dib *et al.*, 2015). Furthermore, phi29 DNA polymerase favours small sequences for amplification (Norman *et al.*, 2014). This feature may explain the prevalence of plasmids <10kbp in studies using the protocol (Li *et al.*, 2012), which mirrors the size constraints of TRACA (Jones & Marchesi, 2007).

1.4 Computational Methods

Computational methods offer a high-throughput approach that take advantage of standard genetic sequencing outputs to identify plasmid-derived sequences. A variety of methods exist to detect plasmids from sequence data (see Table 1). The methods are varied in purpose and taxonomic specificity, but often fail to find new plasmids because of a heavy reliance on known plasmid sequences in databases. Most methods are restricted to use in single-species isolates of well-characterised microorganisms. In this sense, bioinformatic plasmid discovery suffers from the same constraints as *in vitro* methods: it is fed by and reinforces the bias toward culturable microorganisms. The advancement of plasmid research requires tools that are capable of finding plasmids from metagenomes. This way, uncultured microbes can be included in the analysis to better inform our understanding of plasmid diversity and function.

Table 1. *A collection of computational tools that have been developed for plasmid discovery in sequenced isolates or metagenome data (indicated in column 'Data source'). Special mention is made of tools which are only appropriate for use on extensively-studied bacterial taxa or have been created for use on a specific taxon.*

Plasmid identification and reconstruction tools						
Name	**Year**	**Paper**	**Data source**	**Classify/Reconstruct**	**Database-dependent**	**Method**
cBar	2010	Zhou & Xu, 2010	Isolates & MG[1]	Classify	No	Pentamer frequency analysis
qnr	2012	Boulund *et al.*, 2012	Isolates & MG	Detect specific genes	In training phase	HMM constructed from sequence alignment of fluoroquinolone resistance genes
PlasmidFinder	2014	Carattoli *et al.*, 2014	Isolates*	Classify	Yes	Comparison with replicon database
PLACNET	2014	Lanza *et al.*, 2014	Isolates	Reconstruct	Yes	Visualises contigs against reference plasmid sequences
PlasmidSPAdes	2016	Antipov *et al.*, 2016	Isolates	Classify and reconstruct	No	Finds circular sequences in de Bruijn graph
Recyler	2017	Rozov *et al.*, 2017	Isolates & MG	Classify and reconstruct	No	Coverage-based de Bruijn graph analysis
PlaScope	2018	Royer *et al.*, 2018	Isolates*	Classify	Yes	Based on Centrifuge (Armand *et al.*, 2005) with custom database
PlasmidTRON	2018	Page *et al.*, 2018	Isolates	N/A	No	Associates plasmid gene with observed phenotype
PlasFlow	2018	Krawczyk *et al.*, 2018	Isolates & MG	Classify	No	Deep neural network based on k-mers
MOB-Suite	2018	Robertson & Nash, 2018	Isolates*	Classify and Reconstruct	Yes	Modular suite for clustering, recognition, & typing
PlasmidSeeker	2018	Roosaare *et al.*, 2018	Isolates	Classify	Yes	K-mer raw read analysis
mlPlasmids	2018	Arredondo-Alonso *et al.*, 2018	Isolates*	Classify	In training phase	Trains an ML model using pentamer frequencies from resolved genomes
MetaplasmidSPAdes	2019	Antipov *et al.*, 2019	MG	Reconstruct	No	Finds circular sequences in de Bruijn graph
HyAsP	2019	Muller & Chauve, 2019	Isolates	Reconstruct	Yes	Generates contig chains from assembly graph

[1] MG = Metagenome
* = well-characterised taxa

1.5 Plasmid Discovery in Metagenomes

Metagenomics permits the reconstruction of microbial genomes without the need for cultivation (Tyson *et al.*, 2004). Total environmental DNA is extracted, fragmented, sequenced in short-reads (100-150bp), and assembled into contigs. Algorithms separate assembled contigs into different bins as putative genomes, termed Metagenome-Assembled Genomes (MAGs), based on k-mer profiles (Saeed *et al.*, 2012, Yang *et al.*, 2010), differential read coverage (Nielsen *et al.*, 2014, Wu & Ye, 2011), or a mixture between these methods (Wu *et al.*, 2014, Alneberg *et al.*, 2014, Albertsen *et al.*, 2013, Kang *et al.*, 2015). Genomes can then be recovered *in silico,* provided that genome-wide coverage is sufficient (Luo *et al.*, 2012).

Bioinformatic tools are now used regularly for *de novo* plasmid discovery from isolates (de Toro *et al.*, 2014, Ojala *et al.*, 2014, Holt *et al.*, 2015, Peter *et al.*, 2017), but much less frequently in metagenomes (Gupta *et al.*, 2018, Tyagi *et al.*, 2019). Only four methods currently exist for metagenomic plasmid discovery: Plasflow, cBar, Recycler, MetaplasmidSPAdes. Moreover, investigations of this type suffer from heterogenous results in which the predicted plasmids are tool-specific (Laczny *et al.*, 2017). This observation is likely driven by divergent methodologies and the specific weaknesses of each.

K-mer based tools (PlasFlow, cBar) classify contigs as chromosomal or plasmid-derived based on their sequence composition. These are heavily database-dependent for the training of the classifier and may not be suitable for poorly-characterised species and environments.

Methods that re-examine the assembly graph for circular paths with elevated coverage (Recyler, MetaplasmidSPAdes) offer a novel approach to plasmid discovery but are foiled by repeats that interfere with assembly so that larger (>50kbp) plasmids are out of reach (Arredondo-Alonso *et al.*, 2017). Repeats are very common in plasmid sequences due to transposition events (Sheppard et al., 2016) and are especially linked with resistance genes (Orlek et al., 2017b). Since circularity-finding methods exploit sequence coverage profiles, they also run into difficulties when met with single-copy plasmids as they cannot be distinguished from the main chromosome by coverage.

Another complication lies in how plasmid discovery from metagenomes severs the link between a plasmid and its host. As all genetic data in a metagenome is fragmented, there is no way to reunite a plasmid with its host. Oligonucleotide frequencies between a plasmid and its host can be similar and %G-C content in a plasmid is usually lower than its host (Bohlin *et al.*, 2008, Nishida, 2012, Rocha & Danchin, 2002). However, using this signature to infer host relationships *in situ* is flawed because the force that shifts a plasmid's genome signature towards that of its host ('fitness cost amelioration') is exerted over successive generations (Dahlberg & Chao, 2003). Recently acquired plasmids show no similarity to the host genome (van Passel *et al.*, 2006) and investigating any such similarity can only give an indication as to the historical (long-term) host of a plasmid (Suzuki *et al.*, 2008). Moreover, the sheer quantity of genomes within a metagenome introduces a lot of complexity into oligonucleotide differentiation which minimises the chance of accurately determining plasmid-host linkages.

Long-read technologies (Nanopore, SMRT) offer enormous potential in identifying plasmids in sequence data due to the long contiguous sequences that they deliver. However, short read sequencing is much cheaper and is by far the most commonly employed sequencing technology. Moreover, there is an enormous trove of short read sequencing data that has been built up over the past decade. Developing innovative methods to find plasmids in ordinary short read data is thus of immense value.

1.6 Rationale for the Recurrent Assembly Method

Given the limitations and variance in results of current methods, there is a need for bioinformatics methods that can identify plasmids in a metagenome without reliance on pre-formed databases.

This thesis describes a novel method for plasmid discovery in metagenomic sequence data called Recurrent Assembly – here manifested in a newly developed tool called RecurM. The foundational idea of the method is that identical assembly of a sequence across multiple metagenomes provides evidence for the integrity of the sequence. If an assembled contig appears many times over, it may constitute a complete, discrete genetic unit. It is hypothesised that plasmids will be among these recurrent elements.

RecurM promises applicability to the vast store of metagenome data that has already been generated because it does not require new laboratory, sequencing, or assembly algorithms to be developed. Furthermore, the analysis of several metagenomes at once constitutes a scalable, whole-environment plasmid survey that is not offered by current methods. It is reference-free, thus agnostic towards pre-existing knowledge of plasmids. Finally, it does not rely on the circularity of plasmids for detection, thus enabling the inclusion of linear plasmids in the analysis. RecurM holds great potential in helping to expand our understanding of plasmid diversity beyond that which is currently known.

1.7 Thawing Permafrost Metagenomes as a Model System for RecurM development

The research described in this thesis was carried out on data from the Permafrost Project at the Australian Centre for Ecogenomics (ACE). Stordalen Mire, a permafrost site in northern Sweden, has been studied for over a decade at ACE, where metagenomics approaches are being used to better characterise the metabolic potential of microbes in the permafrost thaw gradient.

The permafrost is an environment of high concern to scientists. As climate change raises global temperatures, areas of long-frozen permafrost begins to thaw, and large amounts of organic carbon sequestered in the ground are made available for degradation by microbes and conversion into yet more greenhouse gases such as CO_2 and CH_4. Permafrost is estimated to hold more than twice as much carbon as is presently circulating in the atmosphere and thus threatens an enormous positive feedback loop for atmospheric-carbon-induced climate changes (Schuur *et al.*, 2008). To accurately model the future release of greenhouse gases from the thawing permafrost, a more thorough understanding of the microbial ecology of this environment is needed.

Mobile Genetic Elements (MGEs), of which plasmids are a principle component, are an often-overlooked dynamic in microbial ecology. With greater appreciation for the functions, diversity, dissemination and prevalence of plasmids in the permafrost microbiome, models can be improved to better inform our predictions of greenhouse gas release from this critical environment.

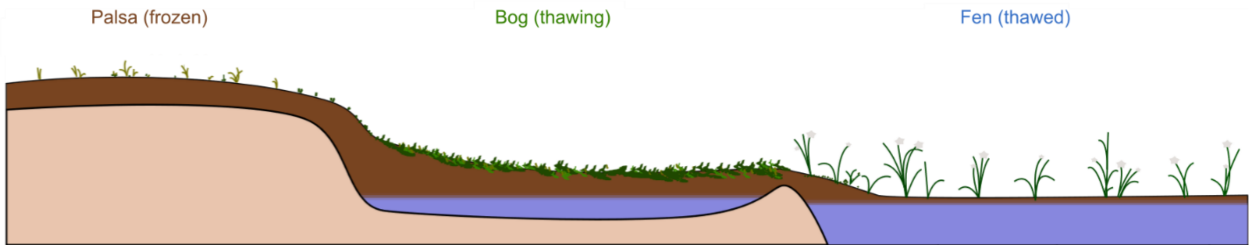

Figure 1. Representation of the permafrost thaw gradient*. Palsa (left) thaws progressively to create bog (centre), and fully saturated fen (right) environments. Thawing is associated with greater activity of microbial methanogens, which degrade organic matter in the bog and fen and release methane into the atmosphere. Samples are taken at different depths of the active layer (brown) along the surface of the gradient. Figure adapted from Woodcroft et al., 2018.*

1.8 Hypothesis & Aims

This project deploys and investigates a recurrent assembly approach for the identification of plasmids in metagenomes. It is hypothesized that sequences repeatedly assembled in different samples will include complete plasmids. To explore the hypothesis, three aims are defined:

i. To develop and program an algorithm for the detection of repeatedly assembled contigs across metagenomes.
ii. To demonstrate the enrichment of plasmid-related genes in repeatedly assembled contigs from ACE's repository of metagenomes from the permafrost.
iii. To compare the recurrent assembly method with existing metagenomic plasmid discovery tools for the recovery of putative plasmids from the permafrost.

Section 2 - Materials and Methods

2.1 Data

All data used in this project were generated from samples taken during the seasonal thaw periods (June – October) of 2010 to 2017 from the peatland environment of Stordalen Mire, Sweden. Soil samples were collected along the entire breadth of the permafrost thaw gradient (palsa, fen, and bog), and at four different depths: shallow (0-5cm), middle (5-15cm), deep (15-25cm), or extra deep (30-48cm). Samples were collected in triplicate and processed for DNA extraction prior to this study according to the protocol described in (Mondav *et al.*, 2014).

Metagenomic libraries were also generated prior to this study using TruSeq Nano DNA Sample Preparation Kit (Illumina, San Diego, USA) and subsequently sequenced using the Illumina NextSeq (Woodcroft *et al.*, 2018). Assembly of these metagenomes was performed by collaborators at Ohio State University using metaSPAdes (Bankevich *et al.*, 2012). Overall, 377 assembled metagenomes were used for input into the RecurM algorithm.

2.2 RecurM Contig-Matching Workflow

The RecurM algorithm consists of four stages in serial: Assembly Preparation and Alignment, Alignment Thresholding, Cluster generation, and Hierarchical Cluster Graphing. Modules for the algorithm were written in Bash and Python version 3.5.0.

2.2.1 Assembly Preparation and Alignment

This stage represents the most computationally intensive stage of the algorithm. Firstly, each metagenome assembly is filtered by length so that only those sequences longer than 2,000 bp are retained for inclusion in the analysis. Assembled contig names are then modified to associate them with the metagenome that they belong to, such that the new sequence name is of the form *">[ASSEMBLYFILE]__[SEQUENCENAME]".* With this modification, each sequence has a unique name and can be traced back to its original source. Finally, these filtered, modified assemblies are aligned with each other using Nucmer version 4.0 (Marçais *et al.*, 2018). Nucmer was selected for its speed improvements over BLAST (Altschul *et al.*, 1990) and for its added precision in alignment information compared to Mash (Ondov *et al.*, 2016). Nucmer's accuracy in aligning genomes of various arrangements and at different rates of nucleotide substitutions is evaluated in Section 3.1.

2.2.2 Alignment Thresholding

The output from Nucmer is parsed and each individually aligned pair of contigs is identified as a recurrent assembly if it meets a numerical threshold against three metrics:

- the Ratio of Lengths between sequence 1 and sequence 2 (RL), given by:

$$RL = \frac{Length_{Seq\ 1}}{Length_{Seq\ 2}}$$

- the Aligned Ratio (AR), calculated on the longer sequence in the alignment and given by:

$$AR = \frac{Alignment\ length}{Length_{Seq\ 1}}$$

- the ANI of the aligned region:

$$ANI = \frac{\#\ of\ mismatches\ in\ alignment}{Alignment\ length}$$

Combined, these metrics indicate both the genetic similarity of the two sequences, as well as the degree to which their assembly can be considered identical. For the purposes of this study, the cutoff value was conservatively set to 90% and an alignment needed to exceed cutoff in all three metrics in order to be classed as a repeated assembly.

Additionally, a fourth metric, RA_{short} is calculated on those matches which pass the threshold at the ANI level but fail in both RL and AR. RA_{short} is calculated on the shorter of the two sequences in the alignment, and given by:

$$AR_{short} = \frac{Alignment\ length}{Length_{Seq\ 2}}$$

2.2.3 Clustering

A disjoint-set algorithm is implemented to perform efficient single-linkage clustering of recurrent sequences. A cluster is constituted when three or more sequences are linked together by alignment linkages. Because it is a single-linkage clustering method, not all sequences in a cluster will exhibit pairwise alignment to each other at a level that passes the 90% thresholds, but all will be connected to each other via a chain of significant alignments. In each cluster, the individual sequence with the greatest number of direct alignments is selected as the representative sequence for that cluster and henceforth termed a Repeatedly Assembled Element (RAE). If two or more sequences share the highest number of direct alignments, one is chosen at random as the representative sequence.

2.2.4 Hierarchical Graphing of Clusters

Clusters may not truly represent discrete RAEs. There may exist alternate assemblies for the same sequences in other clusters. To accommodate this, clusters are transformed into a directed, acyclic, hierarchical graph. Alignments that failed LR, AR and ANI but passed the 90% threshold for AR_{short} are used in this stage as they represent fragmented assemblies. In these fragment alignments, >90% of a shorter sequence's length aligns to <90% of a longer sequence. This functions as a pointer for the shorter sequence towards a longer, alternate assembly of itself.

In the graph, cluster objects are represented as vertices, and edges are drawn using fragment alignments. Fragment alignments were trimmed down to a set wherein both sequences of an alignment were accounted for in clusters, and at least one of the sequences was the representative sequence of its respective cluster. Fragments were then used to graph clusters hierarchically (Figure 2).

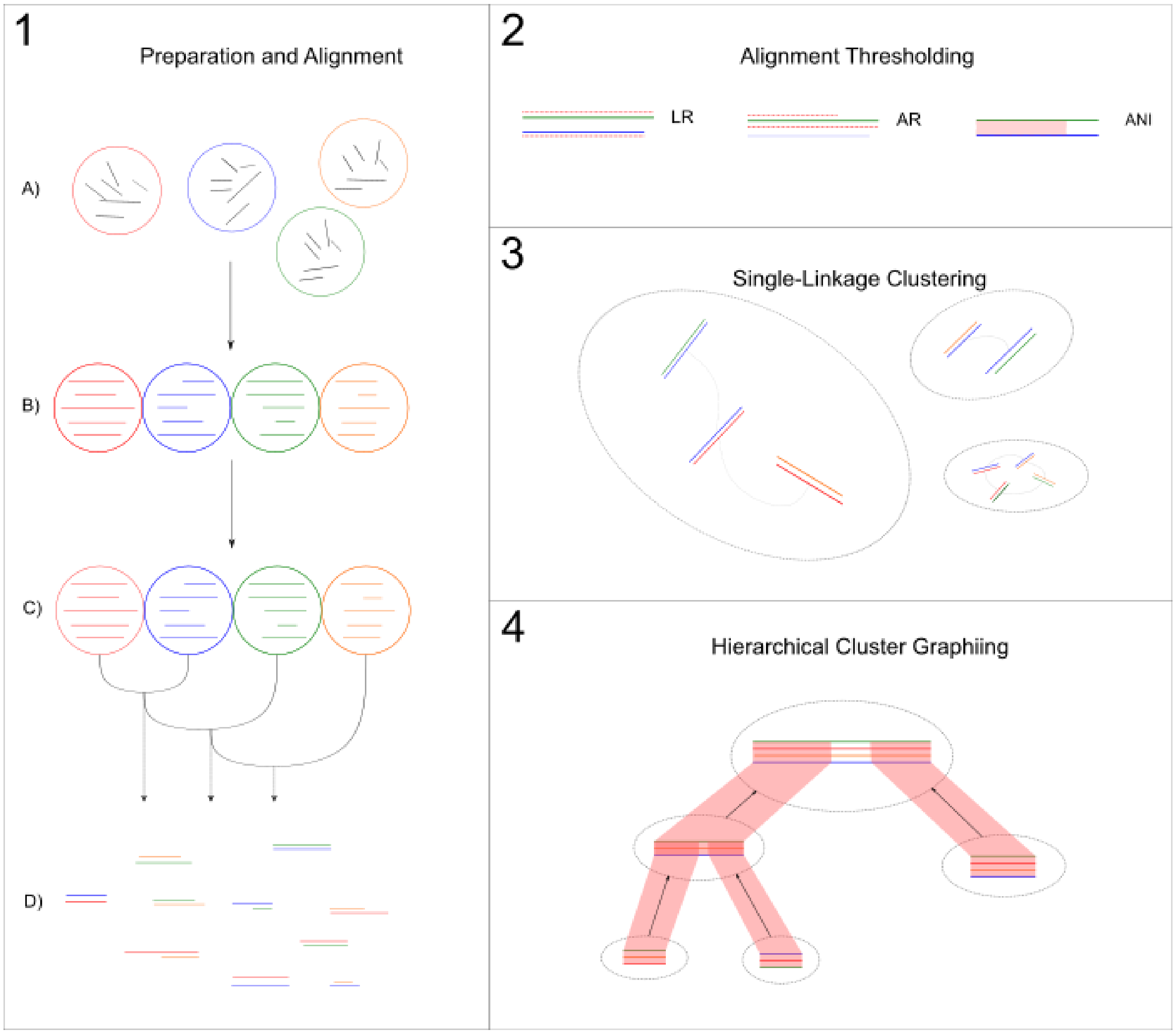
1
Preparation and Alignment
A)
B)
C)
D)
2
Alignment Thresholding
LR
AR
ANI
3
Single-Linkage Clustering
4
Hierarchical Cluster Graphiing

Figure 2. Schematic of the RecurM workflow. 1) *input metagenomes (A) are filtered to a minimum contig length of 2000bp and sequence names are amended so that they include reference to the assembly they originate from (B). Amended metagenomes are aligned with each other in a progressive fashion (C) so that after each Nucmer run the query genome is added to the reference genome to create a reference file that steadily expands through each alignment. This procedure eliminates aligning of metagenomes to themselves and prevents the realigning any pair of metagenomes in the reverse order. As each alignment is completed, raw sequence alignments are added to a pool of alignments (D) for thresholding.* ***2)*** *Sequence alignments must exceed a numerical cutoff for three separate metrics: Length Ratio (LR) which measures the closeness of the sequence lengths, Alignment Ratio (AR) which measures the proportion of a sequence that is actually in alignment, and ANI which measures the genetic similarity between the sequences. Red shading is added to highlight the features being compared in each metric.* ***3)*** *Sequences are clustered together through significant alignments shared with other sequences.* ***4)*** *Clusters are drawn on a directed hierarchical graph depicting their longer and shorter alternative assemblies. Red shading indicates alignment of fragments to longer clusters. Arrows show the direction drawn by the edge.*

2.3 Labelling and Trimming of Clusters

Further analysis of clusters was performed to divide them into types. Single linkage for cluster generation is prone to erroneous inclusion of sequences. If each sequence only needs to pass the 90% threshold to any other sequence in the cluster, a cluster can incorporate elements that are quite dissimilar to the central representative sequence through a chain of single linkages that steadily degrade its similarity to the representative sequence. For this reason, a measure of the clustering closeness was included in the analysis. A cluster was labelled as 'tight' if the representative sequence was directly aligned to >90% of the sequences in the cluster. Only these 56% of clusters were further investigated.

Each cluster was labelled as circular or linear based on the orientation of the sequence alignments within it. In some pairwise alignments, the first half of sequence A aligns to the second half of sequence B and the second half of sequence A aligns back to the first half of sequence B. This alignment is interpreted as circular because each sequence essentially spans the endings of the other, thus circularising the linear representation of the sequence. For a cluster to be labelled as circular, just one of the cluster's alignments needed to be circular because linear alignments can still be present in the cluster if a circular sequence is broken in the same location on two sequences.

Clusters were labelled as linear according to the starts and ends of their alignments. If the alignment between two sequences begins and finishes at the very ends (disregarding interruptions in the alignment), then the ends have been assembled identically. This finding, coupled with its recurrent observance, is taken as evidence that the sequence may be a linear genomic element. For a whole cluster to be labelled as linear, more than half of the pairwise matches in a cluster were required to display perfect end alignments.

The cluster graph was analysed to identify clusters that have no further parent cluster. 'Peak' clusters were identified as those that had <10% of its component sequences aligning as a fragment to the representative sequence of any longer cluster.

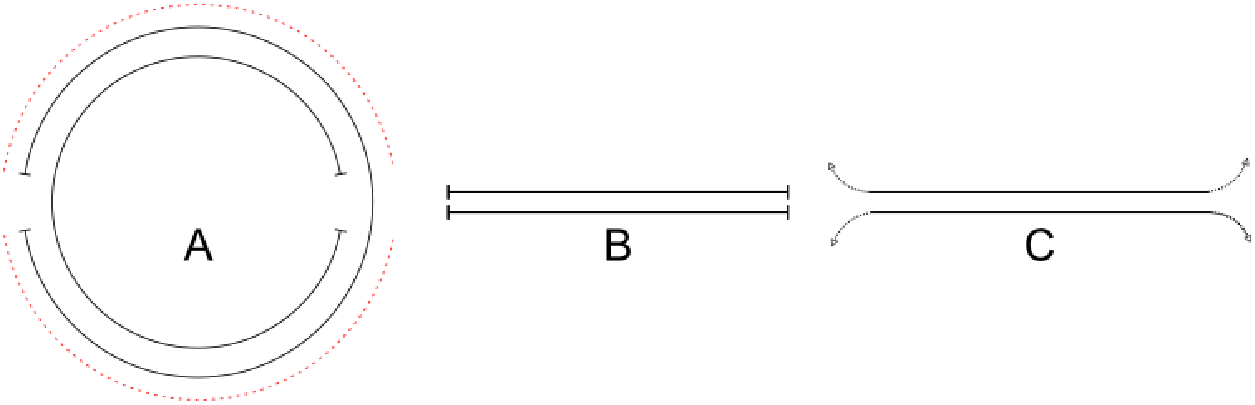

Figure 3. Illustration of the structural labels applied to clusters. A) *Circular arrangement. Two identical sequences but with different break points wrap around each other and resolve the circular structure. Red dashed line indicates the two separate aligned regions seen by Nucmer.* ***B)*** *Linear arrangement. Assembled sequences match perfectly at beginning and end* ***C)*** *Imperfect alignment. Sequences are assembled similarly but are unaligned at the ends.*

2.4 Comparison of RecurM with Existing Metagenomic Plasmid Discovery Tools

The linear and circular RecurM cluster sets were compared with existing tools to determine overlap in plasmid predictions. Recycler and MetaplasmidSPAdes, which require reassembly of read data, are computationally expensive to run over multiple metagenomes, so this comparison was performed on four samples taken from a single permafrost core, 20120800_S3, at shallow, mid, deep, and extra deep sites. However, it should be noted that RecurM relies on scaling of the algorithm with many metagenomes to generate best results and does not perform well on a reduced set. Allowing comparison of RecurM with alternative methods requires preservation of this central functionality. For this reason, RecurM clusters were generated from all 377 metagenomes, but reduced to linear and circular sets containing RAEs that are found in the four target metagenomes.

Since Recycler (Rozov *et al.*, 2017) and MetaplasmidSPAdes (Antipov *et al.*, 2019) explicitly search for circular plasmids, these methods were chosen for comparison of the RecurM circular set. The four permafrost samples were reassembled with SPAdes in '*metagenome'* mode to re-generate the assembly graphs required for these methods to run. Recycler and MetaplasmidSPAdes were run with default settings and with the Recycler *-k* parameter set to 55. Output sequences were filtered by length to exclude those <2000bp and dereplicated using BLAST, ultimately yielding a unique set of circular sequences recovered from the four metagenomes for each of the two tools. Homology between these sets' sequences and RecurM's circular clusters was determined with BLAST, and alignments spanning >90% of any two sequences were deemed identical sequences. Total numbers of putative plasmids and overlap between sets were calculated and visualised in the VennDiagram package for R (Chen & Boutros, 2011).

For comparison of linear plasmids, the K-mer-based methods cBar and PlasFlow were used since these do not require sequences to be circular for detection. cBar and PlasFlow were applied the four pre-assembled metagenomes filtered by length to 2000bp or greater. Overlap between these methods' predictions and RecurM's linear set was determined by matching sequence IDs since these two methods classify already assembled contigs as RecurM does. As above, total numbers of putative plasmids and overlaps were calculated and visualised with VennDiagram. Plasmid predictions from each tool were further compared by interrogating them for plasmid contents following the methods in Section 2.5 below.

2.5 Examination of Plasmid Contents

Confirming whether a contig is a plasmid challenging due to the immense heterogeneity of plasmids. For this project, a rudimentary plasmid identifier was constructed using frequencies of protein families found within the RefSeq Plasmid database. The protein counts were gathered by Antipov et al. (2019) for use in MetaplasmidSPAdes. All 9,937 plasmids in the RefSeq Plasmid database were downloaded and a plasmid-specific HMM database was constructed based on homology of gene contents to proteins in the PFAM-A database. A randomly selected 10% sample of complete bacterial chromosomes from RefSeq was downloaded and Pfams were counted the same way. For each HMM found in the two sets, a ratio of plasmid-frequency to non-plasmid-frequency was calculated, giving a score of [0-infinity]. Scores near 0 indicate much higher frequency of the PFAM on chromosomes, and scores higher than 1 indicate a higher presence of the PFAM on plasmids.

For each set of clusters being studied in this project, the representative sequences were annotated with Prokka (Seemann, 2014). PFAM IDs for each ORF annotation were extracted and contigs were given an average plasmid score as determined by:

$$SCORE = \frac{\sum_{i}^{i=n} f_i}{n},$$

where *i* denotes an ORF on the contig, *n* is the total number of PFAM annotations, and *f* is the frequency ratio of the PFAM given by the plasmid-specific database described above. PFAM IDs that were not present in the database were ignored. The number of contigs in each category of clusters containing no links to the plasmid database were counted but were otherwise also excluded from the analysis. Results were visualized using the R package ggplot2 (Wickham, 2016).

2.6 Cluster-Host Linkage

A proportional relative abundance approach was used to attempt reassociation of putative plasmids and possible hosts. A set of 630 dereplicated MAGs from Stordalen Mire was previously identified (Woodcroft *et al.*, 2018). All RAEs from Circular and Linear cluster sets were aligned with the 630 MAGs using Nucmer. Some RAEs were already associated with a host MAG by the binning process, so MAG contigs that match to RAEs above RecurM’s 90% threshold were removed from the analysis.

Read data from 230 permafrost samples were individually mapped against MAGs and RAEs to determine relative abundance of these units in each sample. Read coverage calculation was performed using CoverM (Woodcroft *et al,* unpublished) with the ‘*genome’* subcommand and minimap2 (Li, 2018) as the underlying read mapping software. Relative abundance of MAGs and RAEs were analysed for proportionality throughout the samples using the propr package in R (Quinn *et al.*, 2017). Rho was chosen for the proportionality metric and centre-log-ratio used as the transformation method on relative abundance counts. RAE-MAG pairs showing significant proportionality ($\rho > 0.95$) were identified as potential plasmid-host links.

Section 3 – Results

3.1 Development of RecurM Alignment using Nucmer

RecurM depends on faithful interpretation of Nucmer alignments in order to accurately identify repeat assembly. One indicator of reliability is that RecurM's alignment metrics should behave predictably and similarly at varying threshold levels. To test the stability of alignments at different threshold levels, the distribution of contig lengths in Nucmer alignments identified as identical were compared at threshold levels of 90% and 95%.

Distributions of contig lengths are similar in shape, but the distribution becomes more biased towards longer sequences as threshold levels increase (Figure 4). This observation is a consequence of the blanket threshold being applied to sequences of all lengths. With a 90% threshold, for example, a sequence of 2000 bp needs 1800 nucleotides in alignment to pass threshold, but a sequence of length 100kbp needs only 80kb. Consequently, high thresholds will accept long sequences with more absolute divergence than would be tolerated in short sequences. This is a flaw that needs to be considered in future development of RecurM, but for this thesis is ameliorated by focussing primarily on linear and circular cluster arrangements.

RecurM's accuracy also depends on the reliability of Nucmer. For Nucmer to provide accurate alignment results with real data, it must be able to handle nucleotide mismatches due to substitution or sequencing errors to a reasonable level. Nucmer alignment and thresholding was performed on sequences with varying error rates simulated under a random model and with linear and circular genome arrangements, revealing satisfactory performance.

The three alignment threshold statistics (LR, AR, ANI) used by RecurM to identify repeated assembly are accurate at mismatch rates equal to or <20% (Figure 5). Alignment is maintained between two sequences up until this level of mismatches. Above a 20% rate of divergence, alignment length deteriorates and AR falls precipitously. This is inconsequential in the context of RecurM, which in any case will quickly identify sequences with a 20% mismatch rate as too divergent to constitute repeated assembly.

Nucmer's alignment results are unaffected by the input of circular sequences with breakpoints in different sections of the circle. The algorithm is able to align regions locally and the three alignment metrics demonstrate the same pattern as shown in linear sequence alignments.

While the aligned length deteriorates rapidly with a rising error rate, the estimated ANI falls at a much steadier rate because ANI is calculated only over the segments of sequence that are aligned. ANI stability may be problematic for the algorithm because any ANI estimate that only accounts for the most similar regions in an alignment is prone to overestimation of sequence similarity. In RecurM, this issue is minimized because ANI is greater than AR and LR in most potential clusters. Marginal contig alignments are likely to fail AR and LR first before ANI influences the result (Figure 6). Sequences that exhibit a non-random error profile (due to conserved regions, for example) may present an exception to this pattern that requires further investigation.

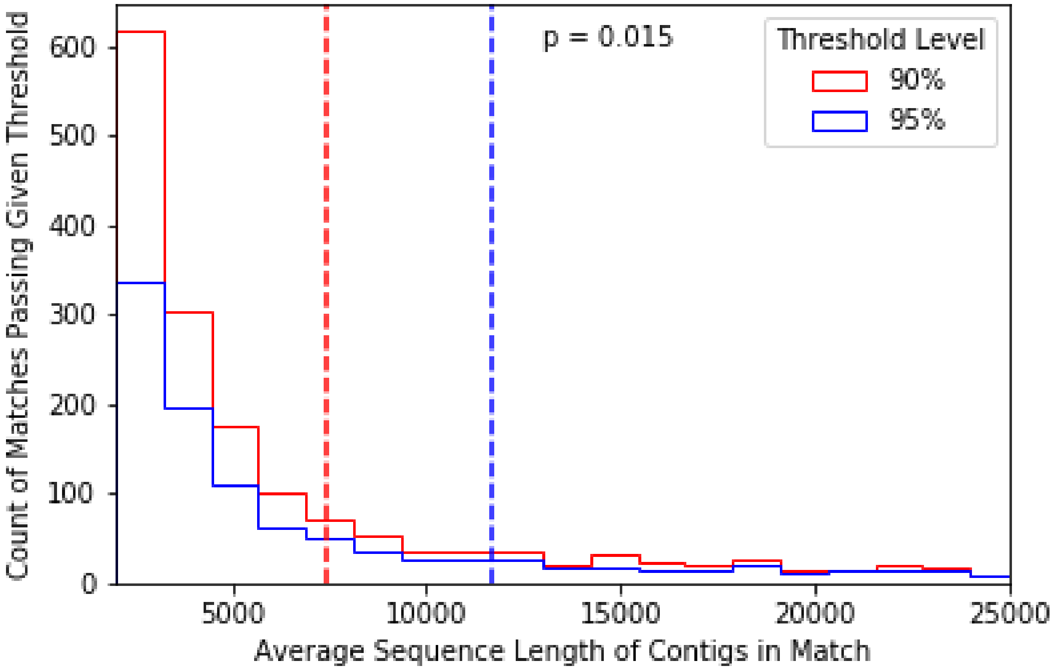

***Figure 4. Length distributions of match objects that pass blanket threshold levels of 95% and 90**%. The mean length of contigs passing threshold at the 95% level is significantly larger (p=0.015) than at 90% according to Welch's two-sample t-test. The increase in length indicates a bias in the thresholding stage towards sequence alignments of longer length at higher threshold levels.*

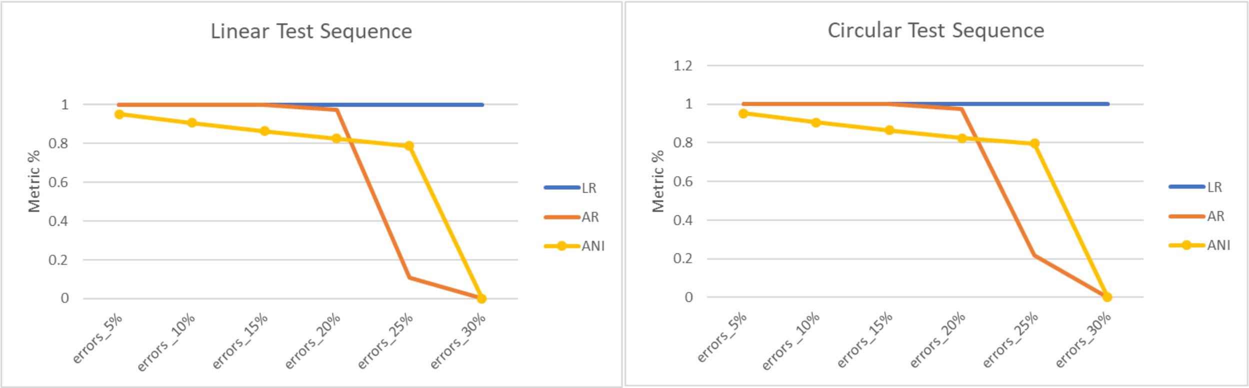

***Figure 5. Performance of alignment metrics under simulated rates of sequence mismatches and under linear (left) and circular (right) arrangements.** AR falls rapidly at an error rate of ~20% as the aligned region collapses. ANI remains stable until alignment fails completely. Metrics are unaffected by arrangement of the genome. Mutations were simulated by selecting x% of bases from one sequence at random (where x is the given divergence rate) and reassigning each base to a randomly chosen alternate nucleotide.*

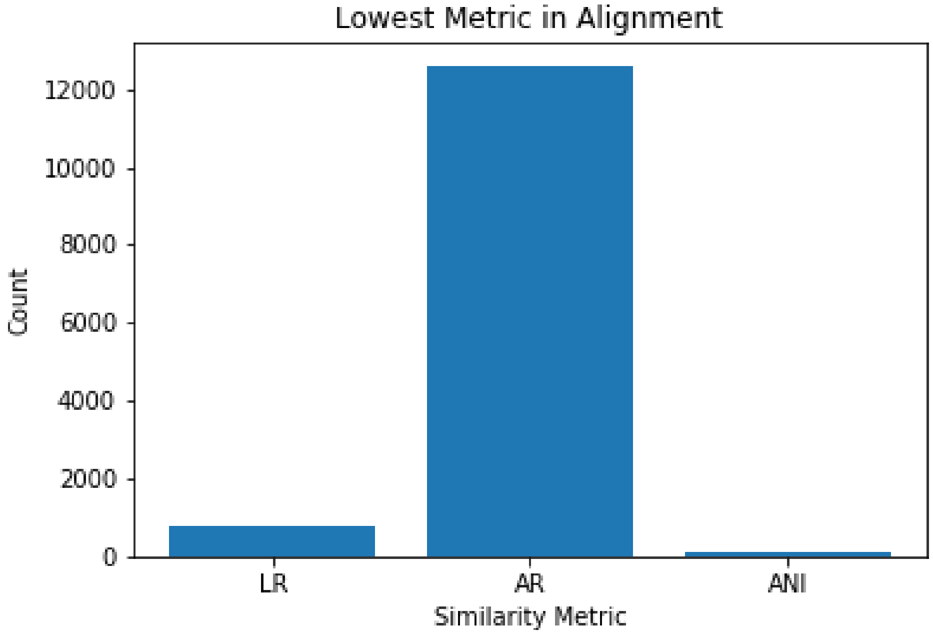

Figure 6. Match objects counted by their lowest Alignment Metric. *Match objects from a four-metagenome RecurM alignment run were initially thresholded to a level of 50% to eliminate the most incomplete alignments. The lowest metric on each of the remaining match objects was then counted. AR is overwhelmingly the most common minimum metric, so is the most likely measurement to trigger failure of a given threshold. The elevated estimation of sequence similarity provided by ANI is of little consequence.*

3.2 Comparison of RecurM with Existing Metagenomic Plasmid Discovery Tools

Quantity of putative circular plasmids discovered with RecurM compares favourably with existing tools (Figure 7). RecurM and Recycler recovered similar quantities (41 & 39 respectively), while MetplasmidSPAdes vastly underperformed them both (11). Overlap of RecurM and Recycler was small, suggesting that their different methodologies has a major impact on the output. The fidelity of the sequences found with Recycler cannot be confirmed, but RecurM's approach depends explicitly on the recurrence of a sequence across metagenomes for it to be counted.

RecurM found far fewer putative linear plasmids than predicted by cBar and PlasFlow. The quantity of plasmids predicted by these alternative tools makes any overlap with RecurM of negligible significance. The high prediction rate from cBar and PlasFlow is suspect, given that the plasmid predictions represent a full 40% and 18% of the total number of contigs in the metagenomes, respectively. It is possible that the high prediction number includes plasmids that are too long for complete assembly and whose fragments have been predicted as plasmids by cBar and PlasFlow, but missed entirely by RecurM. This possibility nonetheless is insufficient to explain the number seen here. These results warrant suspicion and caution around using such methods in lesser-characterised environments like that being studied here.

Comparison of plasmid contents between tools is shown in Figure 8. Against a background of a random sample of contigs from the metagenomes, average contig plasmids scores showed that all tools enrich for plasmid contents. Among tools applied to circular sequences, RecurM predictions showed the highest enrichment, followed by Recycler and MetaplasmidSPAdes. While compelling, reliability of these results is hampered by the small sample sizes included. More than half of contigs from each set were excluded from the analysis due to insufficient PFAM annotations. MetaplasmidSPAdes was plotted with a sample size of only 5.

Conversely, RecurM linear plasmid predictions performed no better than cBar and PlasFlow. No statistical significance is observed in the differences of their means using Welch's two-sample t-test (p > 0.25 in all pairwise comparisons). This suggests that while cBar and PlasFlow are prone to false positives, RecurM does not target linear plasmids with any improved accuracy. Alternatively, the accuracy determined here may be underestimated because of poor representation of linear plasmids in databases. Nevertheless, all 3 tools showed enrichment of plasmid contents above background (p <0.001 for each set).

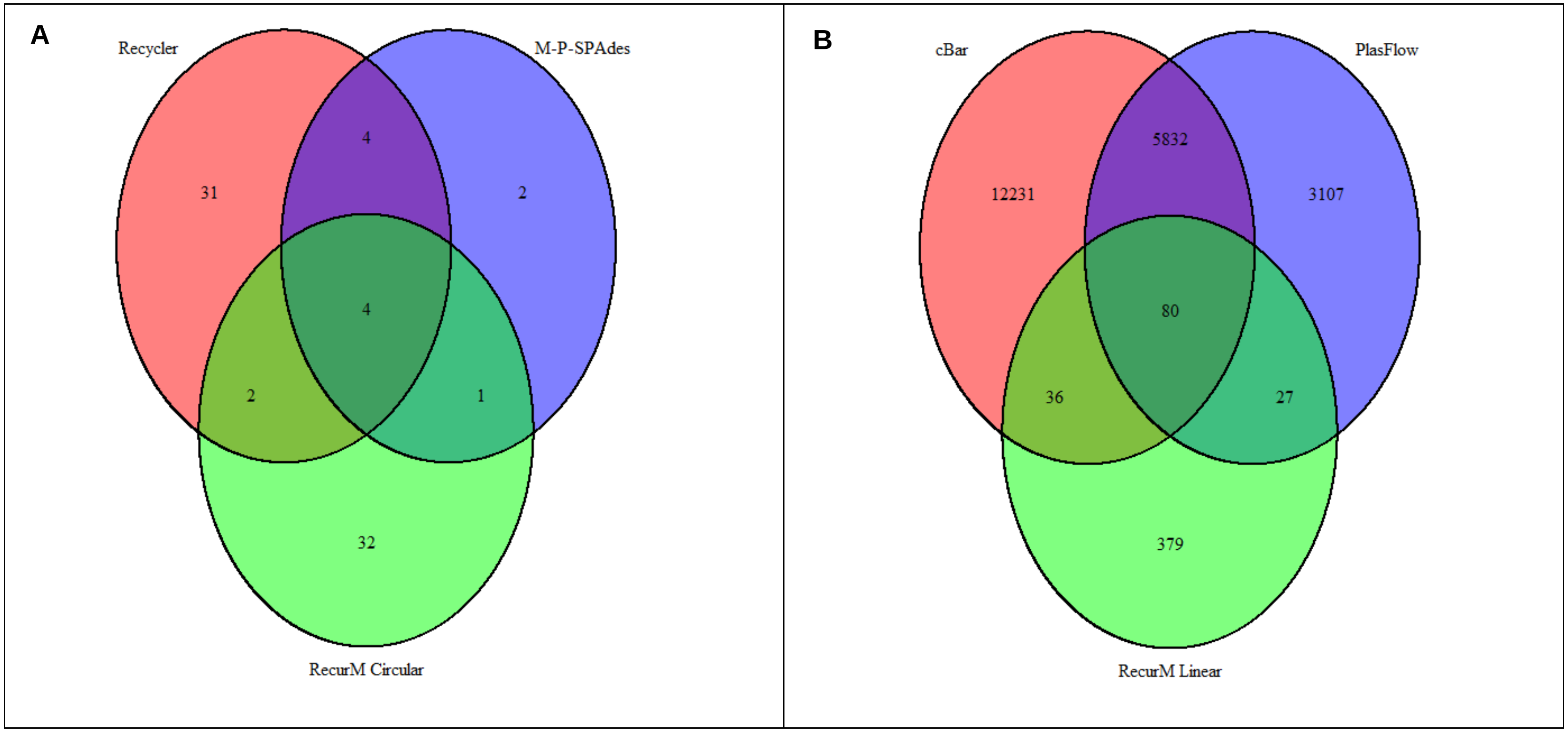

***Figure 7. Venn diagrams of prediction overlap between different plasmid discovery tools**. **A)** Circular sequences detected with Recycler, MetaplasmiSPAdes and RecurM in four metagenomes. **B)** Linear sequences from the same four metagenomes predicted as plasmid by PlasFlow and cBar and identified as linear clusters by RecurM.*

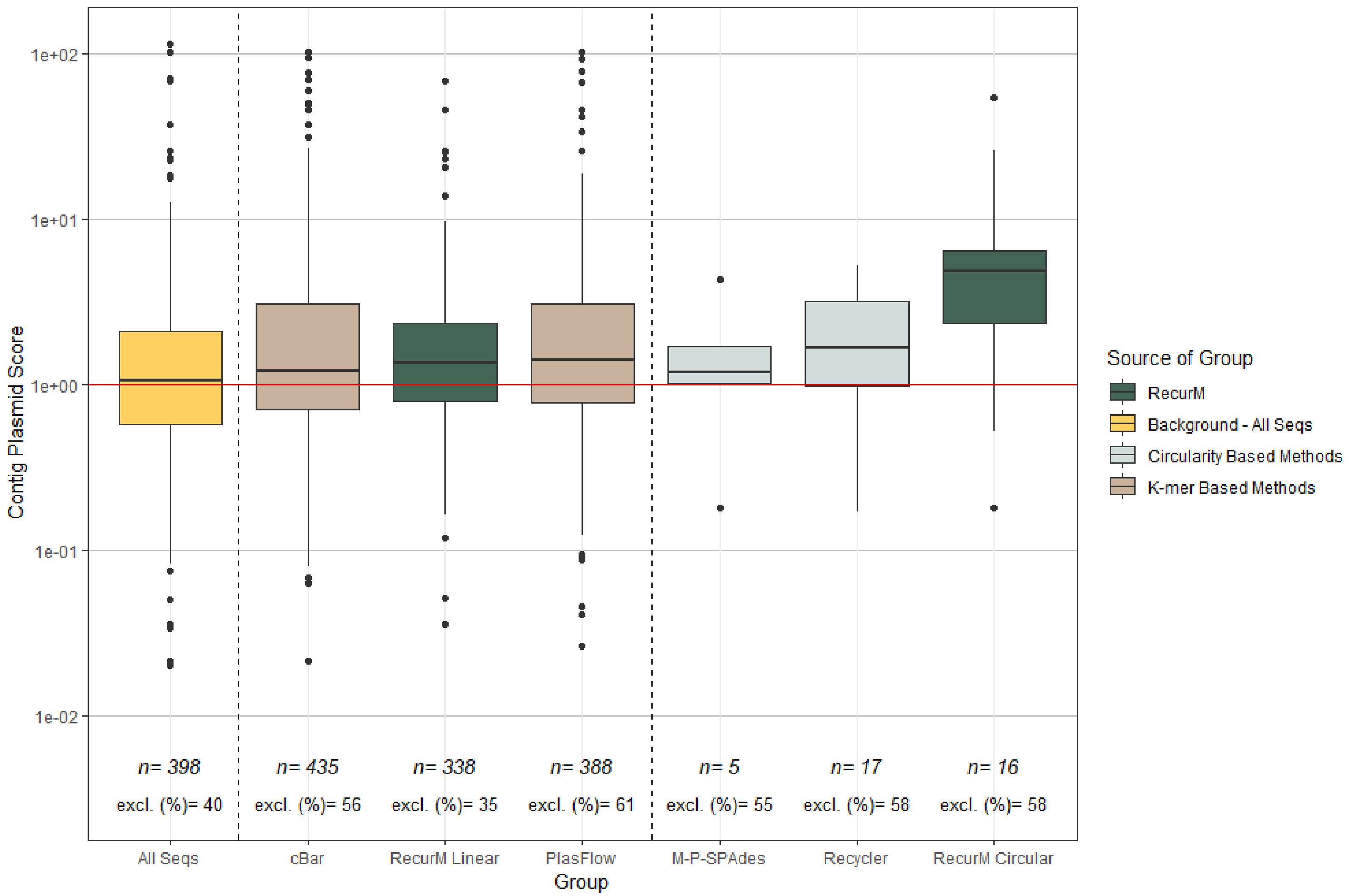

***Figure 8. Boxplot of plasmid scores by plasmid prediction tool**. Dashed vertical lines separate plots by sequence type: background of all sequences (left), linear sequence predictions (middle), and circular sequence predictions (right). Red horizontal line marks the inflection point, above which a sequence score is weighted towards plasmid and below which is weighted towards chromosome. Sample sizes are indicated below each plot along with the fraction of contigs excluded from each set due to insufficient Pfam annotations.*

3.3 Clustering of RAEs from All Metagnomes Enriches for Putative Plasmids

Thresholding of Nucmer output from pairwise alignment of all 377 permafrost metagenomes yielded 4,338,813 unique, significant, similarly assembled alignments of sequences. These alignments were agglomerated into 192,230 clusters. Clusters were filtered to improve quality by applying the following criteria for acceptance: 'peak', meaning <10% of sequences in the cluster align as fragments to a longer cluster in the cluster graph; and 'tight', meaning the representative sequence is directly aligned with >90% of all sequences in the cluster. Constraining of the clusters to just those that are 'tight' lowered the total number to 108,281. The distribution of magnitude and length for these clusters is illustrated in Figure 8. Lastly, clusters were further narrowed to a 'peak' set of 53,862 clusters.

The division of clusters into sets from different stages of the algorithm shows the progression of plasmid enrichment through the RecurM method (Figure 9). Raw clusters achieved no significant elevation of plasmid score, but filtering these by 'peak' and 'tight' criteria produced a small but significant increase in plasmid contents against a background sample of all sequences according to Welch's two-sample t-test ($p< 0.005$). The most dramatic increase in plasmid scores resulted from the addition of linear and circular structural labels.

Circular clusters show the greatest increase in average plasmid score. Linear clusters likewise show a statistically significant elevation of plasmid contents, although at a much lower level. More investigation is needed to determine if these sequences are truly linear or rather occur because of consistent breakage of the sequence at the same point, perhaps due to conserved repeat elements.

Overall, the enrichment of plasmid-related material in Linear and Circular RAEs by RecurM is a strong indication that plasmid structures can be isolated effectively using a recurrent assembly approach.

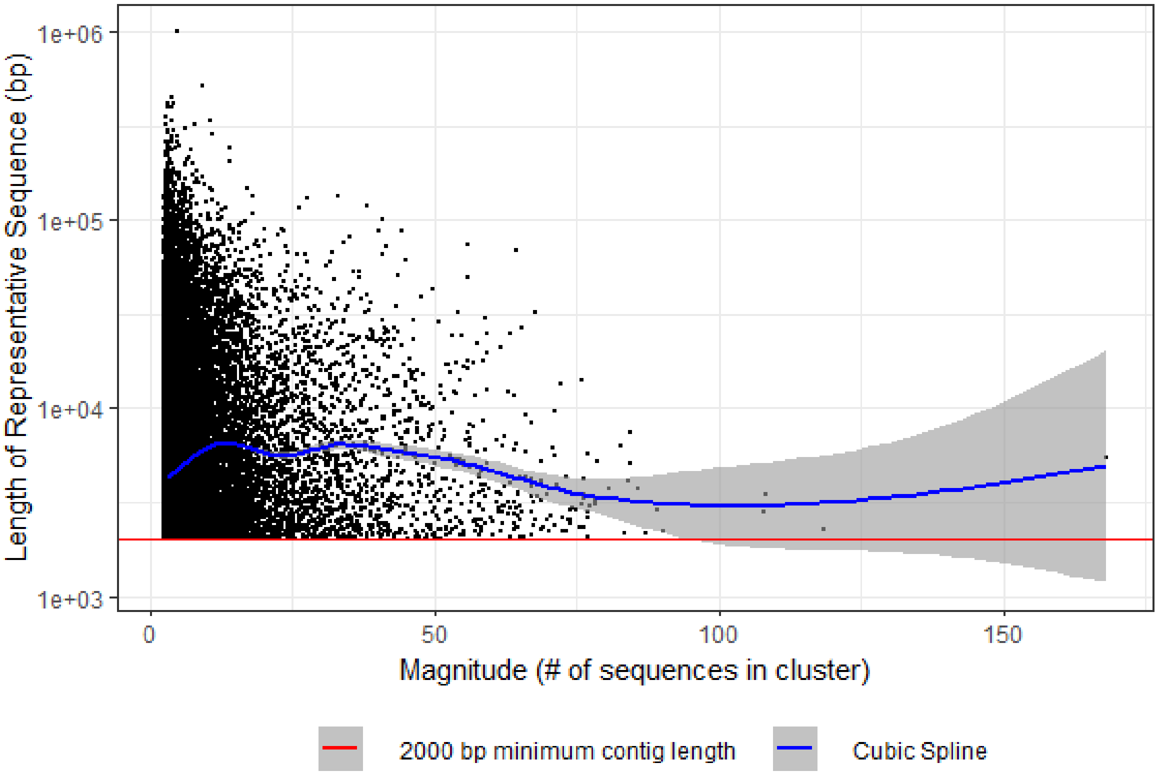

Figure 9. Distribution by length and magnitude of Clusters*. Most clusters are between 5 and 10kb long regardless of how many sequences they contain. There is an elevation in sequence length for clusters of magnitude 10-30.*

***Table 2. Description and summary of cluster sets analyzed for plasmid contents**. Random samples of up to 1000 RAEs were taken from raw clusters (before tightening and cluster graphing), linear clusters and circular clusters, along with a random sampling of 1000 sequences from the metagenomes as a background set. If fewer than 1000 RAEs were present in a category, all components of that category were used in analysis. Some sequences are excluded from analysis because they have no ORFs or contain only hypothetical proteins. Sequences with no ORFs appearing in the plasmid PFAM database are also excluded. A programming error resulted in some sequences being overwritten during extraction, hence some samples contain fewer than the intended 1000 sequences.*

Cluster Set	Description	Total number in set	n (sample)	n excluded for no identifiable ORFs (% of total)	n excluded for no plasmid PFAMs (% of total)	Final number in plot
All Seqs	Random sample of sequences from input metagenomes	~	1000	148 (15)	453 (45)	399
Raw clusters	Random sample of all clusters immediately following clustering step	192,230	998	140 (14)	447 (45)	411
Filtered clusters	Random sample of clusters after filtering to ‘tight’ and ‘peak’	53862	989	112	423	454
Linear < 5kb	Random sample of linear clusters with lengths less than 5000 bp	1958	999	208 (20)	412 (41)	379
Linear > 5kb	Random sample of linear clusters with lengths greater than 5000 bp	1854	1000	12 (1.2)	217 (22)	771
Circular < 5kb	All circular clusters with length less than 5000 bp	292	292	157 (54)	39 (13)	96
Circular > 5kb	All circular clusters with length greater than 5000 bp	163	163	23 (14)	30 (18)	110

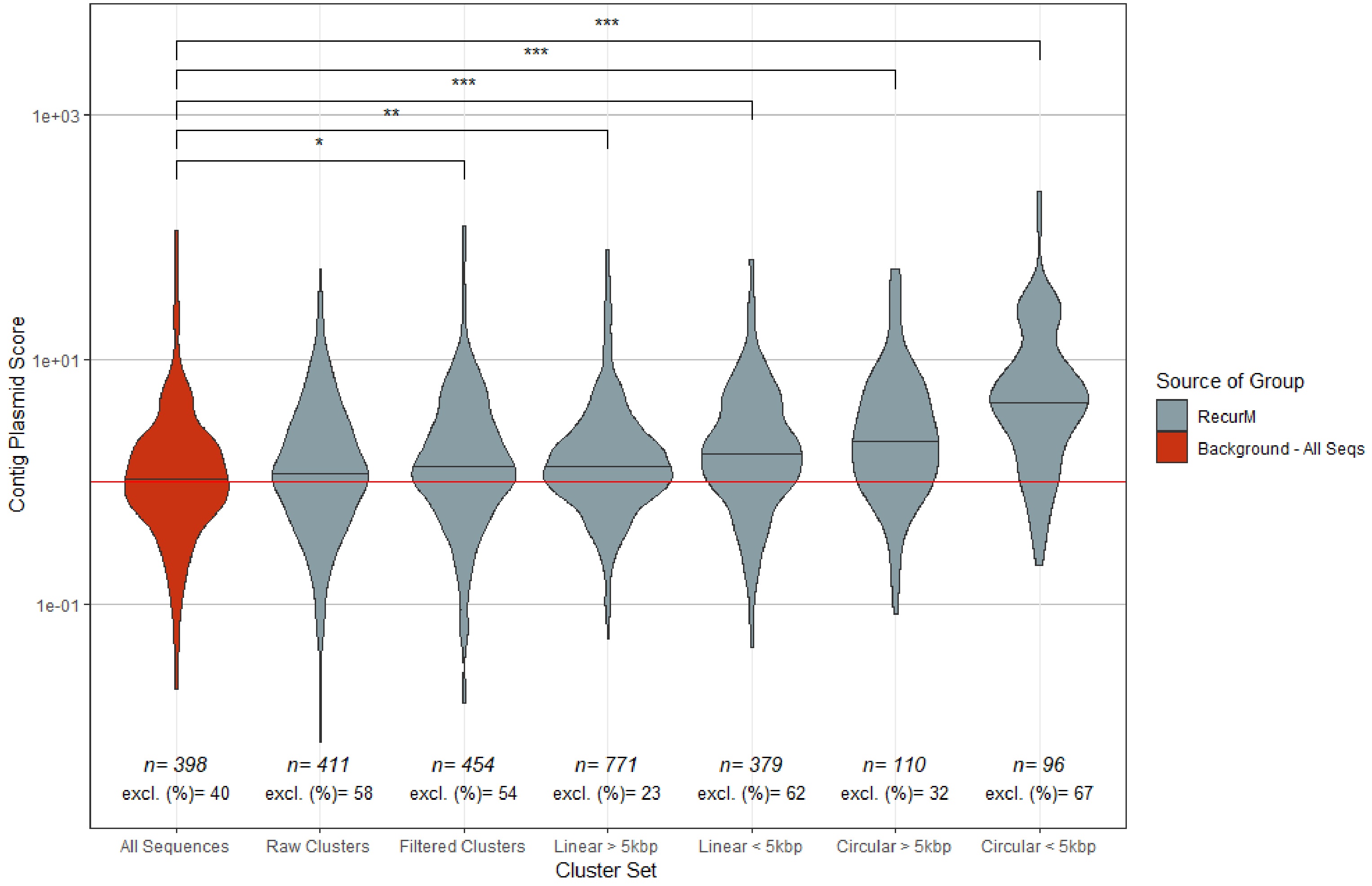

Figure 10. Plot of plasmid score distributions per set of clusters. *Summaries and descriptions of cluster types are found in Table 2. Clusters are arranged left to right by increasing mean, with black bars within plots indicating sample median. Red horizontal line marks the inflection point, above which a sequence score is weighted towards plasmid and below which is weighted towards chromosome. A random sampling of all metagenome samples (Red) is used as background distribution with which to calculate plasmid enrichment in other sets using two-sample Welch's t-test. Star annotations indicate statistical significance between sets at magnitudes of p<0.05(*), p<1e-3(**) and p< 1e-5(***). The fraction of contigs from each set excluded from the analysis due to insufficient annotation is indicated below each plot.*

3.4 Putative Plasmids Inferred from RAEs

Given the present evidence that circular and linear RAEs are enriched for plasmid contents, clusters were categorised into putative plasmids ranks according the strength of their plasmid scores. RAEs with a plasmid score greater than five were named likely plasmids. RAEs with a score between 1 and 5 were named possible plasmids. Those with a score lower than 1 were labelled unlikely plasmid due to the presence of generally conserved chromosomal elements. Clusters with no recognisable features or with no Pfams represented in the database were set aside as unknown clusters.

Counts of these putative plasmids are summarised in Table 3 and their distributions illustrated in Figure 10. CC3, a circular cluster of magnitude 3 and length 14,207bp, is one clear example of a likely plasmid. CC3 contains two chromosome partitioning proteins (ParA and Smc) and a Replicase family protein, which are highly conserved proteins for plasmid replication and dissemination. These features, combined with the circularity of the RAE, provide compelling evidence that the cluster represents a recurrent plasmid.

Table 3. Summary of plasmid categories derived from circular and linear clusters*. Unknown clusters consist of sequences with no known annotations and with annotations unrecognized the plasmid score database.*

Set	**Total**	**Likely Plasmids (Score > 5)**	**Possible Plasmids (1 < Score < 5)**	**Unlikely Plasmids (Score<1)**	**Unknown Clusters**
Circular	455	71	96	35	253
Linear	3812	260	1210	723	1619

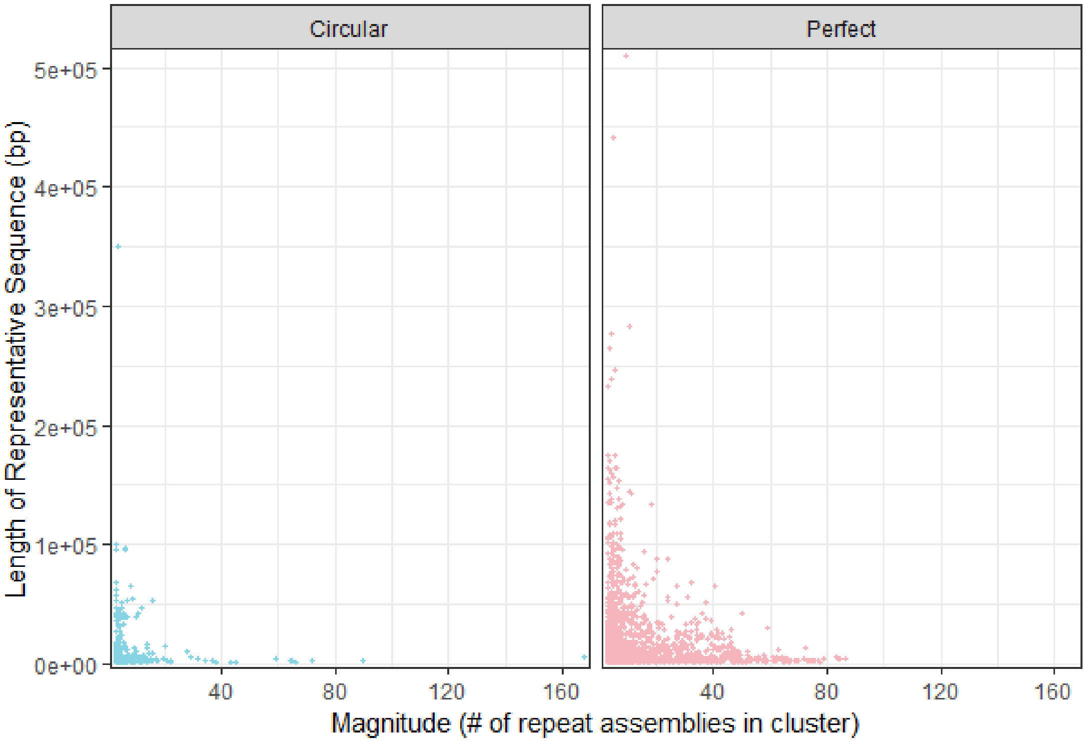

Figure 11. Circular and Linear clusters represented by their magnitude and length of the representative sequence*. Longer sequences generally recur fewer times and vice versa.*

3.5 Differential Coverage Analysis Across Multiple Samples Identifies Potential Hosts of RAEs

All 4,267 putative plasmids (PPs) identified by RecurM were subjected to a proportionality test to link them with MAGs. 947 RAEs were already present in the 630 MAGs, 936 of which were linear and 11 were circular. Three linear elements were found in multiple MAGs. All 947 RAEs found in MAGs were removed from further analysis due to a higher likelihood that these are chromosomal elements. The remaining 3,314 RAEs were analysed with the 630 MAGs for coverage relationships as described in Section 2.5.

In total, 466 significant relative abundance correlations were found between PPs and MAGs ($\rho > 0.95$): 52 circular and 414 linear. Intriguingly, 262 exhibited unique 1-to-1 relationships with MAGs, and 94 showed associations with multiple MAGs. No two PPs exhibited significant co-abundance with each other, indicating that none are likely to be present together in the one microogranism.

By examining the closeness of relationship between a PP and MAG, a hypothesis can be made as to the natural host of the RAE. However, this approach is complicated by the frequent correlation of PPs with more than 1 MAG.

For example, one circular cluster (CC1) of length 47,279 bp demonstrates proportional abundance with 2 MAGs: PALSA-1104 (GenBank ID 6563668), of class *Polyangia*, and AV80 (GenBank ID 6639118), of class *Verrucomicrobiae* (Figure 11). These genomes also share near-significant proportional abundance with each other ($\rho = 0.925$), so incidental correlation for one of this cluster’s linkages cannot be ruled out and it is unclear which genomes, if not both, the PP best shares an association with.

CC1 constitutes a probable plasmid. It lacks any definitive plasmid genes but contains a Recombination endonuclease VII (PF02945) and a putative metallopeptidase domain (PF13203). These two genes are responsible for most of the elevation in score for this contig. Many hypothetical proteins are also present (59). As is true for many of the putative plasmids discovered here, this is an example of where the lack of diversity in databases impacts on the ability to determine whether an element is a plasmid.

In contrast, a circular cluster (CC2) of magnitude 3 with a representative sequence length of 95,227 base pairs displayed a strong 1-to-1 proportional relationship (ρ = 0.988) with a 4Mbp MAG of phylum *Acidobacteria* (Figure 12). In addition to the circular label from RecurM, the representative sequence of CC2 itself was observed to circularise when examined for homology at the sequence ends. Segments of length 500bp from each end of the assembled contig were BLAST to each other and a matching 27 base pair overlap was found, indicating a circular sequence that has been arranged into a linear sequence during assembly. The other two sequences in the cluster shared this finding with overlaps of up to 67 bp.

The most conserved plasmid features are a Domain of Unknown Function (DUF932) and Endonuclease VI (PF02945). CC2 also contains a DNA Polymerase I and a predicted chromosome segregation protein. These features are suggestive of a discrete genetic unit with its own DNA maintenance system such as a plasmid. A phage could also fit this description, but it has no clearly identifiable viral domains and the viral prediction tool Virsorter (Roux *et al.*, 2015) did not classify it is as such. A BLAST search against the NCBI nt/nr database returned a single non-significant (e-value = 2.3) 28bp hit.

Using GTDB-tk (Parks *et al.*, 2018) to assign taxonomy, the correlated MAG was phylogenetically placed in a clade with *Candidatus Sulfotelmatobacter* (GenBank ID 6275958), a draft sulphur-degrading species described only one year ago (Hausmann *et al.*, 2018). The full 5.39Mbp draft genome of this novel species was downloaded and BLAST was used to check for presence of CC2 in it. No hits were returned, indicating that the draft genome had not previously been associated with CC2.

If the association identified by proportionality analysis here is correct, CC2 would be the first extrachromosomal element identified in *Candidatus Sulfotelmatobacter*. More investigation is required to confirm the association, however, as the appearance of the genome in only three metagenomes is too rare to prompt a high degree of confidence in the association. Indeed, sparsity of data is a major problem in metagenomics which frequently undermines confidence in detected correlations (Tsilimigras & Fodor, 2016).

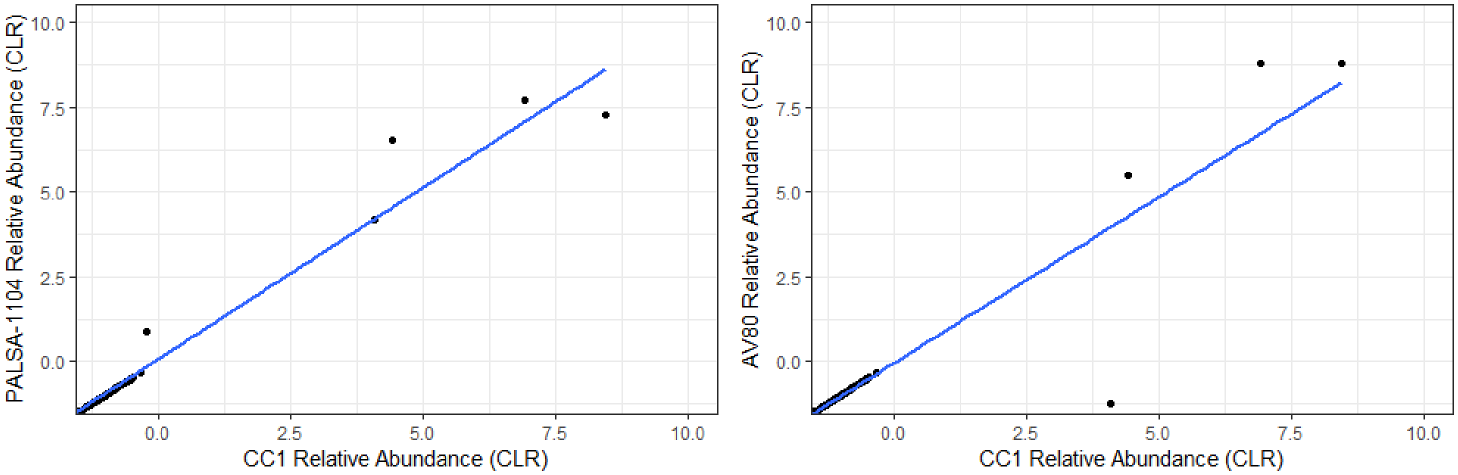

***Figure 12. Correlated relative abundance between CC1 and 2 MAGs**: PALSA-1104 and AV80. Each point represents a centre-log transformed relative abundance of the putative plasmid and genome, after a sample-dependent pseudocount has been added. A linear relationship is shown, with zero relative abundance points included to demonstrate anchoring of the relationship on these points.*

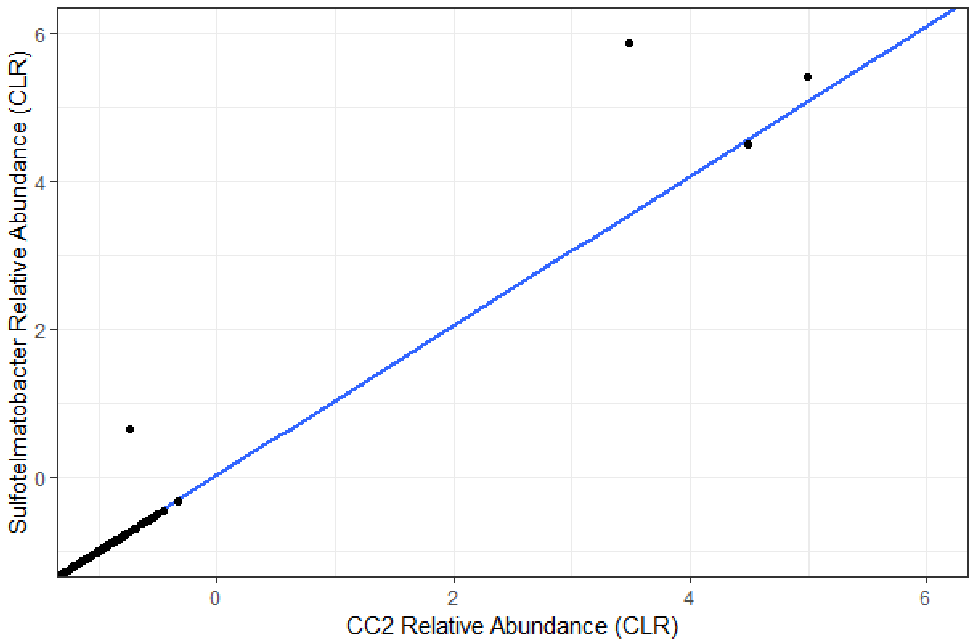

***Figure 13. Correlated relative abundance between CC2 and Sulfotelmatobacter spp.** Each point represents a centre-log transformed relative abundance of the putative plasmid and genome, after a sample-dependent pseudocount has been added. A linear regression is shown. The point with positive Sulfotelmatobacter spp. abundance and negative CC2 abundance indicates a non-zero relative abundance for the genome and zero relative abundance estimated for the plasmid. The point may represent plasticity in Sulfotelmatobacter spp. plasmid content across samples or may be an artefact of the regularisation procedure used in the relative abundance estimation software CoverM.*

Section 4 - Discussion

This thesis presents the conceptualisation and validation of a tool based on Recurrent Assembly. RecurM finds plasmids by identifying sequences assembled identically or near-identically across several metagenomes. It requires only assembled metagenomes as input and is therefore reference-free, differentiating it from database-dependent tools (cBar, PlasFLow, PlasmidFinder), and tools that require more complex data such as raw reads and assembly graphs (Recycler, MetaplasmidSPAdes). RecurM is scalable, improving in its power to detect repeat elements as more metagenomes are included. While the speed of the sequence alignment process as implemented right now does not scale, implementation of k-mer based approaches could further reduce CPU time.

The alignment and orientation of identically assembled sequences can be used to infer circularity of the assembled unit, which heightens the probability that the unit may indeed be a plasmid. It is demonstrated here that these repeatedly assembled circular structures are enriched for plasmid-like genes and functions, as the ratio of frequencies for plasmid genes against chromosomal genes is significantly higher in circular clusters (Figure 9). The 71 probable circular plasmids identified here demonstrate proof-of-principle of the recurrent assembly approach.

RecurM's most powerful feature is its disregarding of reference data to find putative plasmids. Most metagenomic reference data are focussed towards clinical microbiology which limits researchers' ability to survey the plasmids in environmental microbiomes. As a manifestation of this, the PFam-based plasmid verifier used here failed when RAEs weren't annotated with ORFs that could be found in the plasmid PFam database. Some contained putative genes not homologous enough to be identified by the HMMs held in PFam, while others contained only hypothetical proteins and no identifiable ORFs at all (25% of all circular RAEs less than 5kbp). Despite their plasmid-like length and clear circularity, these sequences cannot be clearly identified. Past research has revealed that 'Junk' DNA is abundant in some prokaryotic species (Gil & Latorre, 2012) which, owing to heavily repetitive elements, may constitute a significant portion of small, unannotated RAEs.

Even so, the permafrost thaw gradient being studied here is a novel environment with potentially divergent plasmid functions. Notwithstanding a few culture-based studies (Petrova *et al.*, 2014; Kholodii *et al.*, 2004), plasmids have never before been characterised en-masse in the permafrost active layer. Just as metagenomics has recovered unexpectedly diverse microbial and viral genomes, there may be entirely new classes of plasmid genes to be discovered for which existing data is unequipped to confirm or repudiate as plasmid-derived.

Avoiding this shortcoming, the sequence feature that RecurM looks for is *structural*. The fact that these contigs have been repeatedly assembled provides a level of confidence in sequence integrity that has not been used in any other tools. Sequences identified by recurrent assembly are strong candidates for confirmation as plasmids through traditional laboratory methods. PCR can be used to amplify and confirm circular sequences and novel FISH-based methods currently in development at ACE can visualise the location of the sequence in a microbiome. Being scalable and high-throughput, RecurM is ideal for a large-scale survey of the permafrost mobilome. It leverages the enormous breadth and depth of metagenome data from this environment. By complementing RecurM with confirmatory laboratory methods and with investigation of unknown features, a database can be built that is more representative of the permafrost.

Circular structures discovered with RecurM are noteworthy for their plasmid-like length (< 100,000bp in 99% of sequences) and for the very fact that their structure has been resolved by circularisation. Even in the absence of clearly plasmid-derived genes, the discrete structures discovered here are smaller than even the smallest known microbe (McCutcheon & von Dohlen, 2011) and are likely extrachromosomal. Echoing the findings of this study, Jorgensen et al. (2014), in a metamobilome survey of a rat cecum, detected 616 similar circular elements, 26% of which contained no recognisable replicon domains. Of all 616 sequences, 95% were confirmed as circular *in vitro* with PCR.

Discovery of ‘cryptic’ plasmids in metagenomes has thus been demonstrated prior to this study. However, findings of this type have previously resulted from laboratory plasmidome extraction techniques that specifically select for circular sequences (Brown Kav *et al.*, 2012). RecurM’s advantage is that it discovers similar elements with no special laboratory preparations and using only assembled metagenomes as input. Furthermore, the size range (2kbp – 100kbp) recovered by RecurM vastly exceeds the ~20kbp upper limit of plasmids recovered from laboratory preparation.

The presence of discrete linear plasmids in clusters could not be validated over the course of this project. Neglect of linear plasmids in current plasmid research means they are underrepresented in databases and more difficult to validate. Moreover, while RecurM's circular label is applied with a high degree of confidence, the linear label carries ambiguity. Interrogation of three linear PPs over the course of this study found that none were assembled as complete units, owing to the presence of paired sequence reads that connected the linear PP to other contigs in the assembly. Some linear PPs may indeed constitute complete linear structures, but many are likely to be fragmented assemblies of larger plasmids or sections of microbial chromosome with conserved repeat regions triggering recurrent assembly. K-mer analysis may assist in identifying chromosomal fragments (Dubinkina *et al.*, 2016), especially in RAEs that are linked with a MAGs via proportional abundance. Integrated Chromosomal Elements, (ICEs) which are more frequently associated with HGT between distance taxa (Cury *et al.*, 2018), may additionally form a component of repeatedly assembled linear sequences. Interrogation of putative plasmids for integrases, which are encoded in ICEs at a higher frequency, is suggested as a future avenue to subdivide RAEs.

One promising approach to test for the completeness of linear sequences is a read mapping tool that identifies contigs with no outward contig connections. If an assembled sequence is truly linear, it is hypothesised that all sequence reads mapping to the contig should only have read pairs mapped upstream or downstream on the same contig. Further, these sequences should show have Poisson-distributed read coverage along its length as a result of decline in matched read pairs at the ends of the sequence.

The hypothesis defined above remains untested and is vulnerable to several possible pitfalls. Chiefly, the assembled arrangement of linear plasmids in metagenomes is an unknown quantity. To this author's knowledge, linear plasmids have never been identified in a metagenome assembly. Repeat regions shared between chromosomes and linear plasmids may disrupt assembly. Hairpin closures and telomeres, which are known to be present at linear plasmid termini (Bao & Cohen, 2001, Ravin, 2011, Kikuchi *et al.*, 1985), may interfere with sequencing or truncate assembly, preventing recurrent assembly of the contig. Terminal Inverted Repeats – the main sequence feature of telomeres - may cause linking of linear plasmid contigs during assembly and therefore preclude the possibility of finding discrete units with the proposed read mapping tool. Bioinformatic and lab-based methods will need to be developed and applied together to resolve these questions. Future work ought to experimentally inspect the consistency and regularity of assembly of linear plasmids before aiming to deduce their presence within a metagenome.

Section 5 – Conclusion

Plasmids are an essential component of a microbial community’s metabolic potential yet are poorly characterized and frequently overlooked in metagenomic studies. In the first implementation of a recurrent assembly algorithm, this thesis demonstrates the effectiveness of this high-throughput, scalable approach to plasmid discovery in metagenomes. RecurM requires no specialised treatment of input data, identifies circular plasmids, and holds promise in identifying linear plasmids.

Applying RecurM to a comprehensive set of 377 metagenome samples from the permafrost active layer at Stordalen Mire, 455 circular and 3,812 linear putative plasmids were recovered (71 and 260 with high likelihood, respectively). This is the first instance of plasmid discovery from this site. Additionally, the recurrence of putative plasmids across metagenomes enables the association of these units with potential hosts.

With further development of RecurM and refinement of the data it produces, this approach is poised to make a significant contribution to our understanding of Mobile Genetic Elements in the permafrost and other environments – host-associated systems, wastewater treatment, mine site remediation – where microbial ecology is important.

Section 6 - References

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., and Nielsen, P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**: 533-538.

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014) Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**: 1144-1146.

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**: 403-410.

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**: 3380-3387.

Antipov, D., Raiko, M., Lapidus, A., and Pevzner, P.A. (2019) Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Research* **29**: 961-968.

Armand, F., Bucher, P., Jongeneel, C.V., and Farmer, E.E. (2005) Rapid and selective surveillance of Arabidopsis thaliana genome annotations with Centrifuge. *Bioinformatics* **21**: 2906-2908.

Arredondo-Alonso, S., Rogers, M.R.C., Braat, J.C., Verschuuren, T.D., Top, J., Corander, J., Willems, R.J.L., and Schurch, A.C. (2018) mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb Genom* **4**.

Arredondo-Alonso, S., Willems, R.J., van Schaik, W., and Schurch, A.C. (2017) On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* **3**: e000128.

Baker-Austin, C., Wright, M.S., Stepanauskas, R., and McArthur, J.V. (2006) Co-selection of antibiotic and metal resistance. *Trends Microbiol* **14**: 176-182.

Bale, M.J., Day, M.J., and Fry, J.C. (1988) Novel method for studying plasmid transfer in undisturbed river epilithon. *Appl Environ Microbiol* **54**: 2756-2758.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., and Pevzner, P.A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455-477.

Bao, K., and Cohen, S.N. (2001) Terminal proteins essential for the replication of linear plasmids and chromosomes in Streptomyces. *Genes & development* **15**: 1518-1527.

Blaisonneau, J., Nosek, J., and Fukuhara, H. (1999) Linear DNA plasmid pPK2 of Pichia kluyveri: distinction between cytoplasmic and mitochondrial linear plasmids in yeasts. *Yeast* **15**: 781-791.

Bohlin, J., Skjerve, E., and Ussery, D.W. (2008) Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. *BMC Genomics* **9**: 104.

Bohm, M.E., Huptas, C., Krey, V.M., and Scherer, S. (2015) Massive horizontal gene transfer, strictly vertical inheritance and ancient duplications differentially shape the evolution of Bacillus cereus enterotoxin operons hbl, cytK and nhe. *BMC Evol Biol* **15**: 246.

Boulund, F., Johnning, A., Pereira, M.B., Larsson, D.G., and Kristiansson, E. (2012) A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences. *BMC Genomics* **13**: 695.

Branger, C., Ledda, A., Billard-Pomares, T., Doublet, B., Fouteau, S., Barbe, V., Roche, D., Cruveiller, S., Medigue, C., Castellanos, M., Decre, D., Drieux-Rouze, L., Clermont, O., Glodt, J., Tenaillon, O., Cloeckaert, A., Arlet, G., and Denamur, E. (2018) Extended-spectrum beta-lactamase-encoding genes are spreading on a wide range of Escherichia coli plasmids existing prior to the use of third-generation cephalosporins. *Microb Genom* **4**.

Brom, S., Garcia de los Santos, A., Stepkowsky, T., Flores, M., Davila, G., Romero, D., and Palacios, R. (1992) Different plasmids of Rhizobium leguminosarum bv. phaseoli are required for optimal symbiotic performance. *J Bacteriol* **174**: 5183-5189.

Brown Kav, A., Benhar, I., and Mizrahi, I. (2013) A method for purifying high quality and high yield plasmid DNA for metagenomic and deep sequencing approaches. *J Microbiol Methods* **95**: 272-279.

Brown Kav, A., Sasson, G., Jami, E., Doron-Faigenboim, A., Benhar, I., and Mizrahi, I. (2012) Insights into the bovine rumen plasmidome. *Proc Natl Acad Sci U S A* **109**: 5452-5457.

Carattoli, A., Zankari, E., Garcia-Fernandez, A., Voldby Larsen, M., Lund, O., Villa, L., Moller Aarestrup, F., and Hasman, H. (2014) In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* **58**: 3895-3903.

Casjens, S., Palmer, N., Van Vugt, R., Mun Huang, W., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., Dodson, R.J., Haft, D., Hickey, E., Gwinn, M., White, O., and M. Fraser, C. (2002) A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete Borrelia burgdorferi. *Molecular Microbiology* **35**: 490-516.

Chen, H., and Boutros, P.C. (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**: 35.

Conlan, S., Thomas, P.J., Deming, C., Park, M., Lau, A.F., Dekker, J.P., Snitkin, E.S., Clark, T.A., Luong, K., Song, Y., Tsai, Y.C., Boitano, M., Dayal, J., Brooks, S.Y., Schmidt, B., Young, A.C., Thomas, J.W., Bouffard, G.G., Blakesley, R.W., Program, N.C.S., Mullikin, J.C., Korlach, J., Henderson, D.K., Frank, K.M., Palmore, T.N., and Segre, J.A. (2014) Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci Transl Med* **6**: 254ra126.

Cury, J., Oliveira, P.H., De La Cruz, F., and Rocha, E.P.C. (2018) Host Range and Genetic Plasticity Explain the Coexistence of Integrative and Extrachromosomal Mobile Genetic Elements. *Molecular Biology and Evolution* **35**: 2230-2239.

Dahlberg, C., and Chao, L. (2003) Amelioration of the cost of conjugative plasmid carriage in Eschericha coli K12. *Genetics* **165**: 1641-1649.

Dahlberg, C., Linberg, C., Torsvik, V.L., and Hermansson, M. (1997) Conjugative plasmids isolated from bacteria in marine environments show various degrees of homology to each other and are not closely related to well-characterized plasmids. *Appl Environ Microbiol* **63**: 4692-4697.

de Been, M., Lanza, V.F., de Toro, M., Scharringa, J., Dohmen, W., Du, Y., Hu, J., Lei, Y., Li, N., Tooming-Klunderud, A., Heederik, D.J., Fluit, A.C., Bonten, M.J., Willems, R.J., de la Cruz, F., and van Schaik, W. (2014) Dissemination of cephalosporin resistance genes between Escherichia coli strains from farm animals and humans by specific plasmid lineages. *PLoS Genet* **10**: e1004776.

de Toro, M., Garcillaon-Barcia, M.P., and De La Cruz, F. (2014) Plasmid Diversity and Adaptation Analyzed by Massive Sequencing of Escherichia coli Plasmids. *Microbiol Spectr* **2**.

del Solar, G., Giraldo, R., Ruiz-Echevarria, M.J., Espinosa, M., and Diaz-Orejas, R. (1998) Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev* **62**: 434-464.

Dib, J.R., Wagenknecht, M., Farias, M.E., and Meinhardt, F. (2015) Strategies and approaches in plasmidome studies-uncovering plasmid diversity disregarding of linear elements? *Front Microbiol* **6**: 463.

Dubinkina, V.B., Ischenko, D.S., Ulyantsev, V.I., Tyakht, A.V., and Alexeev, D.G. (2016) Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC bioinformatics* **17**: 38-38.

Fondi, M., Bacci, G., Brilli, M., Papaleo, M.C., Mengoni, A., Vaneechoutte, M., Dijkshoorn, L., and Fani, R. (2010) Exploring the evolutionary dynamics of plasmids: the Acinetobacter pan-plasmidome. *BMC Evol Biol* **10**: 59.

Fournes, F., Val, M.E., Skovgaard, O., and Mazel, D. (2018) Replicate Once Per Cell Cycle: Replication Control of Secondary Chromosomes. *Front Microbiol* **9**: 1833.

Garcia-Fernandez, A., Chiaretto, G., Bertini, A., Villa, L., Fortini, D., Ricci, A., and Carattoli, A. (2008) Multilocus sequence typing of IncI1 plasmids carrying extended-spectrum beta-lactamases in Escherichia coli and Salmonella of human and animal origin. *The Journal of antimicrobial chemotherapy* **61**: 1229-1233.

Garcillan-Barcia, M.P., Francia, M.V., and de la Cruz, F. (2009) The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol Rev* **33**: 657-687.

Gil, R., and Latorre, A. (2012) Factors behind junk DNA in bacteria. *Genes* **3**: 634-650.

Gotz, A., Pukall, R., Smit, E., Tietze, E., Prager, R., Tschape, H., van Elsas, J.D., and Smalla, K. (1996) Detection and characterization of broad-host-range plasmids in environmental bacteria by PCR. *Appl Environ Microbiol* **62**: 2621-2628.

Guerra, B., Soto, S., Helmuth, R., and Mendoza, M.C. (2002) Characterization of a self-transferable plasmid from Salmonella enterica serotype typhimurium clinical isolates carrying two integron-borne gene cassettes together with virulence and drug resistance genes. *Antimicrob Agents Chemother* **46**: 2977-2981.

Gunge, N., Murata, K., and Sakaguchi, K. (1982) Transformation of Saccharomyces cerevisiae with linear DNA killer plasmids from Kluyveromyces lactis. *J Bacteriol* **151**: 462-464.

Gupta, S.K., Shin, H., Han, D., Hur, H.G., and Unno, T. (2018) Metagenomic analysis reveals the prevalence and persistence of antibiotic- and heavy metal-resistance genes in wastewater treatment plant. *J Microbiol* **56**: 408-415.

Harrison, P.W., Lower, R.P., Kim, N.K., and Young, J.P. (2010) Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol* **18**: 141-148.

Hausmann, B., Pelikan, C., Herbold, C.W., Köstlbacher, S., Albertsen, M., Eichorst, S.A., Glavina del Rio, T., Huemer, M., Nielsen, P.H., Rattei, T., Stingl, U., Tringe, S.G., Trojan, D., Wentrup, C., Woebken, D., Pester, M., and Loy, A. (2018) Peatland Acidobacteria with a dissimilatory sulfur metabolism. *The ISME Journal* **12**: 1729-1742.

Hayakawa, T., Otake, N., Yonehara, H., Tanaka, T., and Sakaguchi, K. (1979) Isolation and characterization of plasmids from Streptomyces. *The Journal of antibiotics* **32**: 1348-1350.

Hinnebusch, J., and Tilly, K. (1993) Linear plasmids and chromosomes in bacteria. *Mol Microbiol* **10**: 917-922.

Holt, K.E., Wertheim, H., Zadoks, R.N., Baker, S., Whitehouse, C.A., Dance, D., Jenney, A., Connor, T.R., Hsu, L.Y., Severin, J., Brisse, S., Cao, H., Wilksch, J., Gorrie, C., Schultz, M.B., Edwards, D.J., Nguyen, K.V., Nguyen, T.V., Dao, T.T., Mensink, M., Minh, V.L., Nhu, N.T.K., Schultsz, C., Kuntaman, K., Newton, P.N., Moore, C.E., Strugnell, R.A., and Thomson, N.R. (2015) Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. **112**: E3574-E3581.

Jitwasinkul, T., Suriyaphol, P., Tangphatsornruang, S., Hansen, M.A., Hansen, L.H., Sorensen, S.J., Permpikul, C., Rongrungruang, Y., and Tribuddharat, C. (2016) Plasmid metagenomics reveals multiple antibiotic resistance gene classes among the gut microbiomes of hospitalised patients. *J Glob Antimicrob Resist* **6**: 57-66.

Johnson, T.J., and Nolan, L.K. (2009) Pathogenomics of the virulence plasmids of Escherichia coli. *Microbiol Mol Biol Rev* **73**: 750-774.

Jones, B.V., and Marchesi, J.R. (2007) Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat Methods* **4**: 55-61.

Jorgensen, T.S., Xu, Z., Hansen, M.A., Sorensen, S.J., and Hansen, L.H. (2014) Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metamobilome. *PLoS One* **9**: e87924.

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.

Kholodii, G., Mindlin, S., Gorlenko, Z., Petrova, M., Hobman, J., and Nikiforov, V. (2004) Translocation of transposition-deficient (TndPKLH2-like) transposons in the natural environment: mechanistic insights from the study of adjacent DNA sequences. *Microbiology* **150**: 979-992.

Kikuchi, Y., Hirai, K., Gunge, N., and Hishinuma, F. (1985) Hairpin plasmid--a novel linear DNA of perfect hairpin structure. *The EMBO journal* **4**: 1881-1886.

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

Laczny, C.C., Galata, V., Plum, A., Posch, A.E., and Keller, A. (2017) Assessing the heterogeneity of in silico plasmid predictions based on whole-genome-sequenced clinical isolates. *Brief Bioinform*.

Lanza, V.F., de Toro, M., Garcillan-Barcia, M.P., Mora, A., Blanco, J., Coque, T.M., and de la Cruz, F. (2014) Plasmid flux in Escherichia coli ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genet* **10**: e1004766.

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

Li, L.L., Norman, A., Hansen, L.H., and Sorensen, S.J. (2012) Metamobilomics--expanding our knowledge on the pool of plasmid encoded traits in natural environments using high-throughput sequencing. *Clin Microbiol Infect* **18 Suppl 4**: 5-7.

Lilley, A.K., and Bailey, M.J. (1997) The acquisition of indigenous plasmids by a genetically marked pseudomonad population colonizing the sugar beet phytosphere is related to local environmental conditions. *Appl Environ Microbiol* **63**: 1577-1583.

Lorenzo-Díaz, F., Fernández-López, C., Bravo, A., Espinosa, M., and Lurz, R. (2017) Crosstalk between vertical and horizontal gene transfer: plasmid replication control by a conjugative relaxase. *Nucleic Acids Research* **45**: 7774-7785.

Luo, C., Tsementzi, D., Kyrpides, N.C., and Konstantinidis, K.T. (2012) Individual genome assembly from complex community short-read metagenomic datasets. *ISME J* **6**: 898-901.

Mann, S.P., Hazlewood, G.P., and Orpin, C.G. (1986) Characterization of a cryptic plasmid (p0M1) inButyrivibrio fibrisolvens by restriction endonuclease analysis and its cloning inEscherichia coli. *Current Microbiology* **13**: 17-22.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018) MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology* **14**: e1005944.

McCutcheon, John P., and von Dohlen, Carol D. (2011) An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. *Current Biology* **21**: 1366-1372.

Mondav, R., Woodcroft, B.J., Kim, E.-H., McCalley, C.K., Hodgkins, S.B., Crill, P.M., Chanton, J., Hurst, G.B., VerBerkmoes, N.C., Saleska, S.R., Hugenholtz, P., Rich, V.I., and Tyson, G.W. (2014) Discovery of a novel methanogen prevalent in thawing permafrost. *Nature Communications* **5**: 3212.

Muller, R., and Chauve, C. (2019) HyAsP, a greedy tool for plasmids identification. *Bioinformatics*.

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., Pelletier, E., Bonde, I., Nielsen, T., Manichanh, C., Arumugam, M., Batto, J.M., Quintanilha Dos Santos, M.B., Blom, N., Borruel, N., Burgdorf, K.S., Boumezbeur, F., Casellas, F., Dore, J., Dworzynski, P., Guarner, F., Hansen, T., Hildebrand, F., Kaas, R.S., Kennedy, S., Kristiansen, K., Kultima, J.R., Leonard, P., Levenez, F., Lund, O., Moumen, B., Le Paslier, D., Pons, N., Pedersen, O., Prifti, E., Qin, J., Raes, J., Sorensen, S., Tap, J., Tims, S., Ussery, D.W., Yamada, T., Meta, H.I.T.C., Renault, P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S., Ehrlich, S.D., and Meta, H.I.T.C. (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* **32**: 822-828.

Nishida, H. (2012) Comparative analyses of base compositions, DNA sizes, and dinucleotide frequency profiles in archaeal and bacterial chromosomes and plasmids. *Int J Evol Biol* **2012**: 342482.

Norman, A., Riber, L., Luo, W., Li, L.L., Hansen, L.H., and Sorensen, S.J. (2014) An improved method for including upper size range plasmids in metamobilomes. *PLoS One* **9**: e104405.

Ojala, T., Kankainen, M., Castro, J., Cerca, N., Edelman, S., Westerlund-Wikstrom, B., Paulin, L., Holm, L., and Auvinen, P. (2014) Comparative genomics of Lactobacillus crispatus suggests novel mechanisms for the competitive exclusion of Gardnerella vaginalis. *BMC Genomics* **15**: 1070.

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**: 132.

Orlek, A., Phan, H., Sheppard, A.E., Doumith, M., Ellington, M., Peto, T., Crook, D., Walker, A.S., Woodford, N., Anjum, M.F., and Stoesser, N. (2017a) Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid* **91**: 42-52.

Orlek, A., Stoesser, N., Anjum, M.F., Doumith, M., Ellington, M.J., Peto, T., Crook, D., Woodford, N., Walker, A.S., Phan, H., and Sheppard, A.E. (2017b) Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. *Front Microbiol* **8**: 182.

Page, A.J., Wailan, A., Shao, Y., Judge, K., Dougan, G., Klemm, E.J., Thomson, N.R., and Keane, J.A. (2018) PlasmidTron: assembling the cause of phenotypes and genotypes from NGS data. *Microb Genom* **4**.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **36**: 996.

Peter, S., Oberhettinger, P., Schuele, L., Dinkelacker, A., Vogel, W., Dorfel, D., Bezdan, D., Ossowski, S., Marschal, M., Liese, J., and Willmann, M. (2017) Genomic characterisation of clinical and environmental Pseudomonas putida group strains and determination of their role in the transfer of antimicrobial resistance genes to Pseudomonas aeruginosa. *BMC Genomics* **18**: 859.

Petrova, M., Kurakov, A., Shcherbatova, N., and Mindlin, S. (2014) Genetic structure and biological properties of the first ancient multiresistance plasmid pKLH80 isolated from a permafrost bacterium. *Microbiology* **160**: 2253-2263.

Pistorio, M., Giusti, M.A., Del Papa, M.F., Draghi, W.O., Lozano, M.J., Tejerizo, G.T., and Lagares, A. (2008) Conjugal properties of the Sinorhizobium meliloti plasmid mobilome. *FEMS Microbiol Ecol* **65**: 372-382.

Quinn, T.P., Richardson, M.F., Lovell, D., and Crowley, T.M. (2017) propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci Rep* **7**: 16252.

Ravin, N.V. (2011) N15: the linear phage-plasmid. *Plasmid* **65**: 102-109.

Robertson, J., and Nash, J.H.E. (2018) MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* **4**.

Rocha, E.P.C., and Danchin, A. (2002) Base composition bias might result from competition for metabolic resources. *Trends in Genetics* **18**: 291-294.

Roosaare, M., Puustusmaa, M., Mols, M., Vaher, M., and Remm, M. (2018) PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. *PeerJ* **6**: e4588.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *Peerj* **3**.

Royer, G., Decousser, J.W., Branger, C., Dubois, M., Medigue, C., Denamur, E., and Vallenet, D. (2018) PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom* **4**.

Rozov, R., Brown Kav, A., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., and Shamir, R. (2017) Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* **33**: 475-482.

Saeed, I., Tang, S.L., and Halgamuge, S.K. (2012) Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res* **40**: e34.

Schuur, E.A.G., Bockheim, J., Canadell, J.G., Euskirchen, E., Field, C.B., Goryachkin, S.V., Hagemann, S., Kuhry, P., Lafleur, P.M., Lee, H., Mazhitova, G., Nelson, F.E., Rinke, A., Romanovsky, V.E., Shiklomanov, N., Tarnocai, C., Venevsky, S., Vogel, J.G., and Zimov, S.A. (2008) Vulnerability of permafrost carbon to climate change: Implications for the global carbon cycle. *Bioscience* **58**: 701-714.

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

Segura, A., Molina, L., and Ramos, J.L. (2014) Plasmid-Mediated Tolerance Toward Environmental Pollutants. *Microbiol Spectr* **2**.

Sentchilo, V., Mayer, A.P., Guy, L., Miyazaki, R., Green Tringe, S., Barry, K., Malfatti, S., Goessmann, A., Robinson-Rechavi, M., and van der Meer, J.R. (2013) Community-wide plasmid gene mobilization and selection. *ISME J* **7**: 1173-1186.

Sheppard, A.E., Stoesser, N., Wilson, D.J., Sebra, R., Kasarskis, A., Anson, L.W., Giess, A., Pankhurst, L.J., Vaughan, A., Grim, C.J., Cox, H.L., Yeh, A.J., Modernising Medical Microbiology Informatics, G., Sifri, C.D., Walker, A.S., Peto, T.E., Crook, D.W., and Mathers, A.J. (2016) Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene blaKPC. *Antimicrob Agents Chemother* **60**: 3767-3778.

Shintani, M., Sanchez, Z.K., and Kimbara, K. (2015) Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Front Microbiol* **6**: 242.

Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P., and de la Cruz, F. (2010) Mobility of plasmids. *Microbiol Mol Biol Rev* **74**: 434-452.

Sobecky, P.A., Mincer, T.J., Chang, M.C., and Helinski, D.R. (1997) Plasmids isolated from marine sediment microbial communities contain replication and incompatibility regions unrelated to those of known plasmid groups. *Appl Environ Microbiol* **63**: 888-895.

Suzuki, H., Sota, M., Brown, C.J., and Top, E.M. (2008) Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. *Nucleic Acids Res* **36**: e147.

Tsilimigras, M.C.B., and Fodor, A.A. (2016) Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology* **26**: 330-335.

Tyagi, A., Singh, B., Billekallu Thammegowda, N.K., and Singh, N.K. (2019) Shotgun metagenomics offers novel insights into taxonomic compositions, metabolic pathways and antibiotic resistance genes in fish gut microbiome. *Arch Microbiol*.

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**: 37-43.

van Passel, M.W., Bart, A., Luyf, A.C., van Kampen, A.H., and van der Ende, A. (2006) Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics* **7**: 26.

Wickham, H., (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Woodcroft, B.J., and al, e. (Unpublished) CoverM: Read mapping coverage statistics for metagenomics. *Awaiting Publication*.

Woodcroft, B.J., Singleton, C.M., Boyd, J.A., Evans, P.N., Emerson, J.B., Zayed, A.A.F., Hoelzle, R.D., Lamberton, T.O., McCalley, C.K., Hodgkins, S.B., Wilson, R.M., Purvine, S.O., Nicora, C.D., Li, C., Frolking, S., Chanton, J.P., Crill, P.M., Saleska, S.R., Rich, V.I., and Tyson, G.W. (2018) Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**: 49-54.

Wu, Y.W., Tang, Y.H., Tringe, S.G., Simmons, B.A., and Singer, S.W. (2014) MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**: 26.

Wu, Y.W., and Ye, Y. (2011) A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J Comput Biol* **18**: 523-534.

Yang, B., Peng, Y., Leung, H.C., Yiu, S.M., Chen, J.C., and Chin, F.Y. (2010) Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. *BMC Bioinformatics* **11 Suppl 2**: S5.

Yano, H., Shintani, M., Tomita, M., Suzuki, H., and Oshima, T. (2019) Reconsidering plasmid maintenance factors for computational plasmid design. *Comput Struct Biotechnol J* **17**: 70-81.

Zahran, H.H. (2017) Plasmids impact on rhizobia-legumes symbiosis in diverse environments. *Symbiosis* **73**: 75-91.

Zhou, F., and Xu, Y. (2010) cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* **26**: 2051-2052.