Graph analysis of fragmented long-read bacterial genome assemblies

Pierre Marijon [1,*], Rayan Chikhi [2] and Jean-Stéphane Varré [3]

[1]Inria, Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France.
[2]Institut Pasteur, C3BI USR 3756 IP CNRS, Paris, France
[3]Univ. Lille, CNRS, Centrale Lille, Inria, UMR 9189 - CRIStAL, F-59000 Lille, France.

Abstract

Motivation: Long-read genome assembly tools are expected to reconstruct bacterial genomes nearly perfectly, however they still produce fragmented assemblies in some cases. It would be beneficial to understand whether these cases are intrinsically impossible to resolve, or if assemblers are at fault, implying that genomes could be refined or even finished with little to no additional experimental cost.

Results: We propose a set of computational techniques to assist inspection of fragmented bacterial genome assemblies, through careful analysis of assembly graphs. By finding paths of overlapping raw reads between pairs of contigs, we recover potential short-range connections between contigs that were lost during the assembly process. We show that our procedure recovers 45% of missing contig adjacencies in fragmented `Canu` assemblies, on samples from the NCTC bacterial sequencing project. We also observe that a simple procedure based on enumerating weighted Hamiltonian cycles can suggest likely contig orderings. In our tests, the correct contig order is ranked first in half of the cases and within the top-3 predictions in nearly all evaluated cases, providing a direction for finishing fragmented long-read assemblies.

Availability: `https://gitlab.inria.fr/pmarijon/knot`

Contact: pierre.marijon@inria.fr

1 INTRODUCTION

Third-generation DNA sequencing using PacBio and Oxford Nanopore instruments is increasingly becoming a go-to technology for constructing reference genomes of non-model prokaryotes and eukaryotes. Longer sequencing reads allow in principle to overcome the reconstruction problems posed by genomic repetitions (Bresler *et al.*, 2013). Direct assembly of second-generation (Illumina) sequencing data typically also results in high consensus accuracy yet generally more fragmented bacterial assemblies (Bankevich *et al.*, 2012). The large-scale ongoing NCTC project aims to assemble and make publicly available 3,000 bacterial strains sequenced using PacBio[1].

Recent works have demonstrated single-contig long-read assemblies of bacterial chromosomes (Koren and Phillippy, 2015; Loman *et al.*, 2015). Therefore, it is natural to ask whether genome assembly is now a solved problem with long reads[2], at minimum for smaller genomes such as bacteria. It turns out that in several cases, bacterial assemblies remain fragmented into a handful of contigs, even with long-read sequencing and recent assembly techniques. Deciding whether an assembly instance is resolved is not always clear due to the presence of plasmids, contaminants and unplaced low-quality reads. In this work, an assembly is considered to be *resolved* if the number of contigs classified as chromosomal is equal to the expected number of chromosomes (generally just one, in the bacterial case).

To date, the NCTC project contains 1,735 samples for which 1,136 have been assembled by the consortium, and among these, 599 (34%) are unresolved according to the criteria above (as in Feb 2019). Later in this article, we will see that even when using multiple recent tools, many assemblies remain fragmented. Therefore there is a clear and unmet need for an investigation that determines whether those samples are intrinsically impossible to resolve, or whether current assembly methods are imperfect.

In this article we have selected a subset of NCTC samples (see Results section) and considered the outputs of three recent assemblers: `Canu`, `Miniasm`, and `HINGE`. We observe that instances where the assembly is fragmented can be challenging to further manually elucidate. In general, assemblers produce an assembly graph where nodes are contigs and edges reflect local sequence proximity in the genome (*adjacency*). In fragmented instances, the final assembly graph is sometimes uninformative due to the absence of edges between contigs, hindering further assembly finishing steps. In such cases, it would be tempting to conclude that the assembly is fragmented due to regions of insufficient sequencing coverage, with no way to determine a likely contig order. However, in a number of cases we found that a lack of connectivity can be due to reads that were discarded early in the assembly pipeline. Here we

[1] https://www.sanger.ac.uk/resources/downloads/bacteria/nctc/

[2] See e.g. `https://huit.re/PJMMA_uF`

will show that contig adjacency information can be computationally recovered from the raw data.

To automatically investigate unresolved assemblies and propose directions for refinement, we introduce a set of *in silico* forensics operations for long-read assemblies, and we built a software framework. Our analyses are based solely on information present in the raw sequencing data in addition to the contigs produced by a given assembly tool, and are not biased by any other source, e.g. a closely related reference genome. For validation purposes only and to explain some of our observations, we will align contigs to a ground truth reference when one is available. Our framework is first tested on synthetic data to illustrate a simple case of fragmentation due to heuristics in the `Canu` assembler. We then show on real data that our method helps recover useful adjacency information between contigs.

Going further, we demonstrate how to use this recovered information to provide likely assembly hypotheses using Hamiltonian paths, through a ranked list of contigs orderings. Obtaining a small set of possible orderings between contigs, knowing that the true genome order is likely one of them, can be instrumental to guide further genome finishing steps.

2 RELATED WORKS

Assembly forensics date back to the Sanger era, e.g. with the AMOSvalidate software (Phillippy *et al.*, 2008), which detects mis-assemblies within contigs using multiple sources of information (e.g. read coverage, properly mapped pairs, clipping). Other tools have been introduced for mis-assembly detection in Illumina data (REAPR (Hunt *et al.*, 2013), FRCbam (Vezzi *et al.*, 2012), Pilon (Walker *et al.*, 2014)) and for PacBio data (VALET (Olson *et al.*, 2017)) using similar principles. Completeness of an assembly can be estimated without any reference, using core genes as a proxy metric, e.g. with BUSCO (Simão *et al.*, 2015) or CheckM (Parks *et al.*, 2015) software. Finally, assembly likelihood metrics have been introduced to assess the fit of an assembly to a probabilistic model of sequencing, via re-mapping reads to the assembly (Clark *et al.*, 2013; Rahman and Pachter, 2013; Ghodsi *et al.*, 2013). For a more complete exposition, refer to a recent survey on metagenomics assembly validation (Olson *et al.*, 2017), that also largely applies to isolates.

For bacterial genomes specifically, several pipelines for *assembly finishing* have been developed (Bosi *et al.*, 2015). They usually take as input an assembly obtained with short-read data and align it to one or multiple close reference genomes, in order to find a contig ordering (Kremer *et al.*, 2017). Recent work has examined the cause of assembly fragmentation for seven bacterial genomes sequenced using PacBio sequencing, and rejected the hypothesis that gaps were caused by strong secondary DNA structure (Utturkar *et al.*, 2017). Instead, low coverage and repetitions appear to be the two main factors for contig termination.

To the best of our knowledge, little work has been carried to investigate assemblies based on the graph of assembled contigs or the initial string graph. Noteworthy exceptions are the Bandage software (an assembly graph visualization tool) (Wick *et al.*, 2015), and the `HINGE` assembler that implements automated repeat handling based on the assembly graph (Kamath *et al.*, 2017). We use Bandage extensively in the present work, and will consider datasets where even `HINGE` failed to produce a single-contig assembly.

3 LONG-READ ASSEMBLERS

Several genome assemblers have been developed to process third-generation sequencing data, either stand-alone (Koren *et al.*, 2017; Li, 2016; Kamath *et al.*, 2017; Lin *et al.*, 2016) or in combination with Illumina data (Ye *et al.*, 2016; Antipov *et al.*, 2015; Wick *et al.*, 2017; Zimin *et al.*, 2013). In this work we will focus on three recent stand-alone assemblers, chosen because of their widespread usage (`Canu`), automated graph analysis algorithms (`HINGE`), and speed/modularity (`Miniasm`). However the techniques are likely to be applicable to a broader set of assemblers.

3.1 Description of `Canu`, `Miniasm`, and `HINGE`

The `Canu` (Koren *et al.*, 2017) assembler consists of three major steps: correction, trimming and contig creation. The first two steps should not be regarded as innocuous pre-processing steps, as they significantly impact the rest of the assembly process. The correction step uses MHAP to perform all-against-all read mapping then generates consensus reads with the falcon_sense tool (Chin *et al.*, 2016). `Canu` then performs overlapping of error-corrected reads with a legacy algorithm from the Celera assembler, named *ovl*. The trimming step detects hairpins, chimeric reads, and low-support regions and subsequently cuts reads. A 'unitigging' step is performed using `bogart`, a modified version of CABOG (Miller *et al.*, 2008), to produce a graph that records only the longest overlaps between corrected reads (termed BOG for 'Best Overlap Graph'). `Canu` generates contigs from this graph and improves their consensus accuracy by re-mapping all reads

The `Miniasm` pipeline consists of two separate tools: `Minimap2` and `Miniasm` (Li, 2016). `Minimap2` finds overlaps between raw reads and outputs alignments. `Miniasm` trims low-coverage regions of reads, then constructs a string graph from `Minimap2` alignments that are suffix-prefix overlaps. `Miniasm` performs simplification on the graph inspired by short-read assembly: transitive reduction, tip removal, bubble popping, and short overlaps removal based on a relative length threshold. After simplifications, non-branching paths are returned as contigs.

The `HINGE` (Kamath *et al.*, 2017) assembler uses raw uncorrected reads (similarly to `Miniasm`) to construct an overlap graph similar to the BOG of `Canu`. `HINGE` attempts to output finished bacterial assemblies through improved repeat-resolution. In cases where there subsist repetitions that are not spanned by reads, `HINGE` provides a visualization of the resulting assembly graph for manual inspection.

3.2 Assembly graphs

Short-read and long-read assemblers output final assembly sequences in FASTA format, and an increasing number of tools also output an assembly graph in Graphical Fragment Assembly (GFA) format[3]. A final long-read **assembly graph** typically consists of all contig sequences as nodes, and a set of overlaps between contigs as

[3] `https://github.com/GFA-spec/GFA-spec`

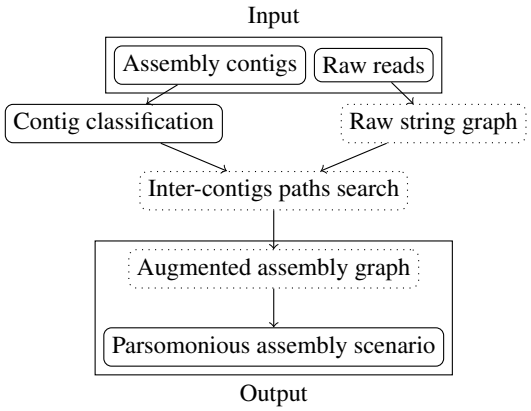

Fig. 1: The proposed framework takes as input raw long-read sequencing data and the output of an assembler. The (optional) contig classification step removes non-chromosomal contigs. A string graph of raw reads is constructed, in which paths are searched between extremities of contigs, then are converted into links between contigs in an augmented assembly graph. When such a graph is connected, putative contig orderings are reported. Dotted nodes represent elements that are automatically visualized in the HTML report.

edges. Assembly graphs are seldom used by downstream tools, and are generally provided for the purpose of inspecting the assembly.

Most long-read assemblers start by constructing then analyzing a *string graph* (**SG**) of the reads (Myers, 2005), where each read is a node, and overlaps between reads are represented by edges to which additional information is attached (e.g. overlap length, overlap error rate). In addition, transitive reduction is performed on the edges and reads that are fully contained in others are discarded.

4 METHODS

We hypothesized that the final contig graph produced by assemblers does not always reflect all the information present in the raw data, and may be missing overlaps or even genomic regions. We built a novel algorithmic framework to recover some of the 'missing' information and further analyze it. The main steps are presented in Fig. 1, and the next sections describe them in more details.

4.1 Raw string graph

First, we eliminate chimeric reads from the raw data based on overlaps found by `Minimap2` using a custom tool[4] (manuscript in preparation (Marijon *et al.*, 2019), see Supplemental Fig. 6). A string graph (SG) is then constructed using overlaps between chimera-removed reads (here, overlaps found by `Minimap2`), considering only the subset of reads not fully mapping inside contigs, i.e. just reads mapping to contig extremities (see Supplemental Section Appendix 1). A stand-alone script was created to convert overlaps from the PAF format (defined in Li (2016)) to a graph in the GFA format[5]. Transitive reduction over the edges of this SG is performed using Myers' algorithm (Myers, 2005).

4.2 Contigs classification

In order to simplify analyses and focus on chromosomal contigs, we filter out contigs of plasmid origin and contigs of unknown taxonomic status (see Supplemental Methods Appendix 2). Contigs that were not marked as chromosomal are discarded. Note however that this contig classification step can be skipped in order to perform analysis of complete, unfiltered sets of contigs.

4.3 Computation of paths between contigs

An essential algorithmic component of our framework is the search for paths in the SG that uncover new connections between contigs. First, one read per contig extremity is identified among reads included in the SG: a read is selected such that both its incoming and outgoing neighbors also map at the same contig extremity (in order to avoid selecting dead-end nodes in the SG).

Then for each pair of contigs, shortest paths between reads at both extremities of each contig are computed in the SG using Dijkstra's algorithm. The length of a path is computed in nucleotides as follows: the sum of all reads lengths involved in the path minus all the overlaps between reads, as well as minus the overlaps between reads and contig extremities. If contigs overlap, the path length is reported as zero. Since we perform path search starting from each contig extremity, we may obtain two shortest paths for each pair of contigs, and only the shortest of those two is kept.

4.4 Augmented assembly graph

We transform a contig graph into a novel object, the *augmented assembly graph* (**AAG**), as follows. Nodes of the AAG are contig extremities. An edge is inserted between two nodes if a path has been found by the procedure in Section 4.3 between the two contig extremities. Each edge is weighted by the corresponding path length. Additionally, zero-weight edges are created between both extremities of each contig.

Such a graph allows to explore adjacencies between contigs, beyond those present in the original contig graph, in order to formulate hypotheses regarding the ordering of contigs. At a certain contig extremity, and in absence of genomic repeats, low-weight edges likely reflect adjacent contigs, while high-weight edges likely correspond to SG paths that pass through other contig(s) (i.e. transitively redundant edges in the AAG). In the presence of repeats, low-weight edges do not necessarily show true adjacencies between contigs, as the true path may be longer. Yet one can observe that a path longer than the longest repeat in the genome necessarily reveals a distant link between two contigs (i.e. necessarily contigs which are truly non-adjacent on the genome), and also such path may go through another contig.

According to Treangen *et al.* (2009) most repetitions in bacteria are shorter than 10kbp. We thus categorize edges of the AAG into 3 groups according to their weight. Consider the path in the SG that led to the creation of the edge e in the AAG between extremities of two different contigs a and b. If the path is longer than 10kbp, and/or

[4] `https://gitlab.inria.fr/pmarijon/yacrd`

[5] `https://gitlab.inria.fr/pmarijon/fpa`

it contains at least one read that was involved in the construction of another contig c, the edge e is named *distant*. Otherwise the edge e is considered to reflect an adjacency between a and b. If there is more than one edge outgoing from the extremity of a or of b, the edge e is named a *multiple adjacency* (likely revealing a putative repeat). Otherwise it is named a *single adjacency*.

4.5 Searching for parsimonious assembly scenarios

We sought to determine whether contigs could possibly be ordered directly using the AAG. In principle, we anticipate to recover a large number of distant edges in the AAG, therefore it would be non-trivial to determine a contig order by direct inspection of the graph layout (e.g. see Fig 3). Given a connected AAG, our working hypothesis is that a minimum-weight Hamiltonian cycle may correspond to the correct contig order (note that having a connected AAG is a necessary condition for such a cycle to exist, but not a sufficient one). This is guided by the intuition that edges in the AAG with high weight are more likely to correspond to false connections due to repetitions or true paths between distant contigs. For simplicity, we search for Hamiltonian cycles and not paths, under the assumption that the genome is circular. We further require that any Hamiltonian cycle traverses all zero-weight edges corresponding to both extremities of each contig. Moreover, contigs mapping inside another one are not considered.

We designed an automated procedure to test this hypothesis, based on computing and sorting Hamiltonian cycles according to their total edge weights. In practice some of the AAGs that we obtain are too complex, due to the presence of short contigs (see the Discussion section for more details). Our pipeline excluded contigs shorter than 100kbp from the AAG before listing all Hamiltonian cycles. For validation purposes, when a reference genome is available, we mapped all chromosomal contigs against this reference to determine the true contig order. We then recovered the position of the true contig order within the list of orders given by Hamiltonian cycles.

4.6 Assembly report generation

We implemented a Snakemake (Koster and Rahmann, 2012) pipeline that takes as input raw reads, contigs produced by an assembler, and optionally a contig graph. The pipeline follows steps described Fig.1, then generates an HTML report for easy inspection. Companion tools to compute AAG edge classification and to perform Hamiltonian path search are also provided.

5 RESULTS

5.1 Datasets

In order to illustrate our methods using a simple yet non-trivial case of assembly graph analysis, we simulated long reads from a linearized reference genome of *Terriglobulus roseus* (NC_018014.1, 5.2 Mbp). This genome contains an unusual 460kbp repeat that is challenging for assembly tools. We used LongISLND (Lau *et al.*, 2016), with 20x sequencing coverage and 9kbp mean read length (Supplemental Table 9).

To investigate real datasets, we mined the NCTC project which consists of 1735 bacterial strains (as of Feb 2019) sequenced using PacBio technology. For each dataset, the NCTC consortium had built an assembly using HGAP and Circlator (Hunt *et al.*, 2015) followed by a manual correction step. We estimate, based on visual inspection of 159 NCTC fragmented `HINGE` assemblies[6] out of 997, that assembly graphs are missing contig adjacency information in 69% of the fragmented assemblies of `HINGE` and `Miniasm`, i.e. around 13% of all NCTC datasets (including those that assemble perfectly). Among datasets for which both `Canu` and `HINGE` failed to produce a single contig per chromosome, we selected 19 datasets where the assembly made by NCTC contains as many chromosomal contigs as the number of expected chromosomes (i.e. is resolved), 24 datasets where the NCTC assembly is unresolved, and finally 2 datasets that were not yet assembled by NCTC. See Supplemental Table 2 for a complete list of the 45 datasets. All datasets were assembled with `Canu` version 1.7 and `Miniasm` version 0.2.

`Canu` contigs were classified according to Section 4.2. On average for each dataset, 10.2% (resp. 6.4%) of the `Canu` (resp. `Miniasm`) contigs are marked as plasmid, 13.7% (resp. 12.2%) do not match any bacteria in the Blast database and are therefore marked as of undefined origin, and the remaining 76.0% (resp. 81.3%) of contigs are classified as chromosomal and are further considered for analysis. Full classification results are presented in Supplemental Table 6 and 7.

We further investigated whether the assemblies could somehow be combined, e.g. by improving `Canu` assemblies using `Miniasm` contigs. We have performed a simple test to evaluate this possibility (see Supplemental section Appendix 3) and could not straightforwardly improve assemblies this way.

5.2 Assembly graph analysis of a synthetic low-coverage dataset

This section gives an introductory overview of the analyses that our method performs on the *T.roseus* simple synthetic dataset described above. `Canu` produced 3 contigs of total length 4.7 Mbp. A ≈500kbp region is missing from the assembly. `Miniasm` produced 7 contigs and the `HINGE` assembler (commit 8613194) was not able to produce an assembly, likely because of the low coverage (20x).

Since the SG has a single connected component (Fig. 2b) but both the BOG and the contig graph of `Canu` have multiple connected components (Fig. 2a), assembly fragmentation can be explained by reads that have been discarded at the BOG construction stage of `Canu`. The coloring of the SG using the connected components of `Canu` BOG (Fig. 2b) further suggests an ordering of contigs. Note that the `Canu` contig graph is uninformative on this dataset, as it contains no edges between contigs.

We performed path analysis as per Section 4.3. Fig. 2d shows the length of paths in SG found between reads at `Canu` contigs extremities. Since a reference genome is available, the true order of contigs is reported on the Figure but note that path analysis does not need this information. We find that the `Canu` contigs named tig8 and tig4 overlap in the SG. tig1 and tig8 are linked by a long path involving 491922bp. This long path can be explained by looking at how tig1 has been built by `Canu`: the path goes through a large 'loop' (see Supplemental Fig. 2) which corresponds to a repeat in the reference (Fig. 2c). The repeat (of length 460kbp) was not resolved by `Canu`, leading to a region of about 440kbp missing

[6] `https://web.stanford.edu/~gkamath/NCTC/report.html`

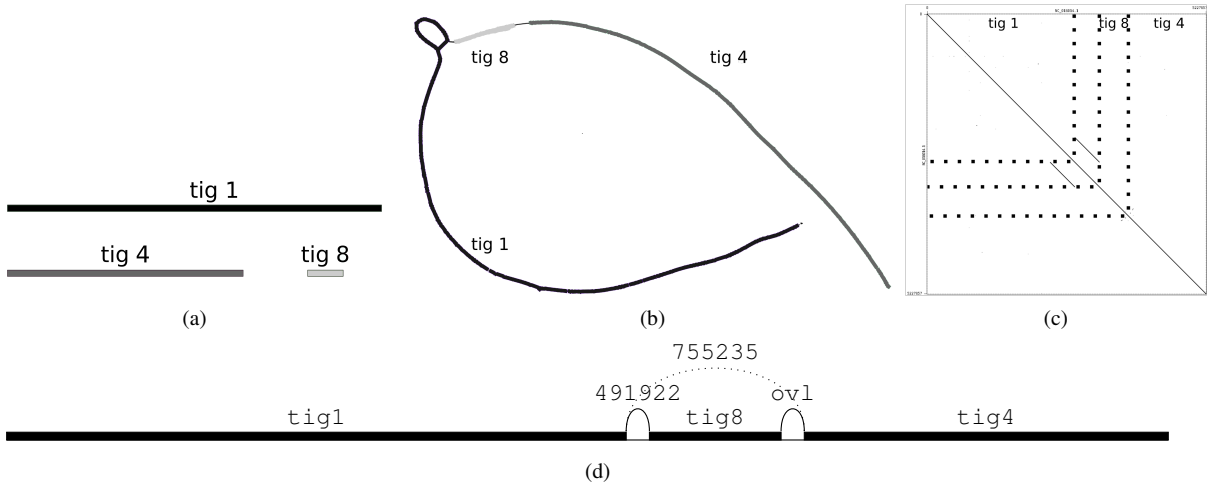

Fig. 2: Graph analysis of a synthetic dataset (*T. roseus*). (a) Contig graph produced by `Canu` (visualized using Bandage): 3 contigs, no edge. (b) SG built from `Minimap2` overlaps, on which connected components of the `Canu` BOG are colored. (c) Dot-plot of the *T. roseus* genome (NC_018014.1) aligned against itself, showing a long tandem repeat. (d) The AAG with `Canu` contigs ordered according to their position on the *T. roseus* reference. If two contigs overlap, no length is given and instead the link is labeled 'ovl'.

from the assembly between tig1 and tig8, which explains why the shortest path between both contigs contains as many as 491922bp. We further checked that the path of length 755235bp between tig1 and tig4 indeed contains reads from tig8, and is therefore redundant. By aligning raw reads and `Canu` corrected reads to the reference genome, we observe a drop of raw reads coverage (around 8x) in the region between tig8 and tig4. This likely explains why `Canu` failed to connect both contigs.

As a side note, a `Canu` assembly of the same dataset with twice higher read coverage (40x) yielded a two-contig assembly, also with same pattern as in between tig8 and tig4. An older version of `Canu` (1.6) fully resolved the 40x dataset into a single contig, likely due to changes in how reads are corrected and trimmed between version 1.6 and 1.7.

5.3 Investigation of 45 unresolved NCTC assemblies

We performed the same type of analysis on the 45 NCTC samples. A `Minimap2` AAG was constructed for each dataset using SG and `Canu` contig extremities. Assembly and AAG statistics are presented in Table 1 for an excerpt of the dataset. Full statistics and more details are given in Supplemental Tables 2, 6 and 7. There we observe that the number of contigs in `Canu` and `Miniasm` assemblies is generally higher than in the assemblies made by NCTC. Nevertheless the sum of lengths of chromosomal contigs is about the same in all assemblies (Supplemental Table 8).

5.3.1 Case study of two NCTC datasets We closely examine two NCTC datasets that contain interesting patterns, through the lens of a ground truth obtained by remapping `Canu` contigs against respective NCTC assemblies using BWA-mem (Li, 2013).

NCTC12123 This dataset was assembled into 5 chromosomal contigs by `Canu`, including 2 contigs that are contained in others and are automatically discarded by our pipeline (see Fig. 3).

The assembly is made of 2 large contigs (tig1 and tig2) and a shorter one (tig9) totaling 4.78 Mbp. `Miniasm` produces also 5 chromosomal contigs, including 3 small ones. Both `Canu` and `Miniasm` contig graphs are made of two components. `HINGE` produces a single-component assembly graph but does not resolve it (because it detects multiple possible traversals). Finally, the NCTC assembly consists of 2 chromosomal contigs: one being 4.69Mbp long and the other 21kbp long. Contigs tig1 and tig2 both map over the large NCTC contig, while tig9 maps to both NCTC contigs. Using the AAG on `Canu` contigs (see Fig. 3), one can observe that a number of scaffolding scenarios could be made following this graph. Interestingly, based on the mapping of the 3 contigs on the larger contig of the NCTC assembly, edges of smaller weight (i.e. shortest paths) tend to be associated with true contig adjacencies. In this example, low-weight Hamiltonian cycles (Section 4.5) yield two likely contig orders (see Supplemental Fig. 3). This SG analysis thus enabled to retrieve an adjacency that was missed by `Canu`. It also confirms the multiple traversals prediction of `HINGE`, further reducing the number of putative contig orders to only two.

NCTC5050 This dataset is assembled into 4 chromosomal contigs by `Canu`, including one that is contained in another. The `Canu` contig graph is 'fully' fragmented as each contig is its own connected component. There is no reference genome for this strain, and we chose as ground truth the NCTC assembly consisting of 2 contigs. One is entirely covered by a `Canu` contig, and the other contains the 3 remaining contigs (see Supplemental Fig. 4). In the following, x_s and x_e denote left (resp. right) extremities of a contig x. We found *single* (i.e. non-repeat) adjacencies between tig1_s/tig23_s, tig1_e/tig10_s, tig10_e/tig23_s that were confirmed by mapping to the longest contig from the NCTC assembly. Together, these single adjacencies suggest a putative scaffolding scenario: tig1 – tig10 – tig23(reversed). This scenario is also the top-ranked one proposed by our Hamiltonian path search procedure (see below).

We also mapped corrected and raw reads to the junction for validation (see Supplemental Fig. 5). We observe a drop of coverage

	NCTC contigs			Canu contigs			# nodes	# dead-ends in		# edges in AAG		
NCTC ID	chr	pls	und	chr	pls	und	in AAG	contig graph	AAG	total	single adjacency	multiple adjacency
NCTC10006	1	0	0	3	0	0	4	2	2	4	2	0
NCTC10332	1	0	0	12	0	0	8	8	4	24	0	3
NCTC10444	1	0	0	7	0	0	8	3	0	24	0	6
NCTC10702	1	1	1	3	3	0	4	4	4	4	0	0
NCTC12123	2	3	0	5	4	1	6	4	1	12	1	2
NCTC12132	1	0	0	2	0	2	4	4	2	4	1	0
NCTC13125	1	2	4	6	3	1	6	0	0	12	0	4
NCTC13463	1	1	4	5	2	2	4	0	0	3	2	0
NCTC5050	2	3	0	4	2	3	6	6	0	12	3	0

Table 1. Assemblies and contig graphs statistics for an excerpt of 9 NCTC datasets (full tables in Supplemental Table 2 and 3), consisting of 8 datasets where Hamiltonian cycle search succeeded, and the NCTC5050 dataset discussed in the Results section. AAGs are constructed using a SG built from Minimap2 overlaps and Canu contig extremities. The 'contig graph' column corresponds to the final assembly graph produced by Canu; 'chr': number of chromosomal contigs; 'pld': number of plasmid contigs; 'und': number of other contigs. Note that some of Canu 'chr' contigs may be contained in others, therefore the '# nodes in AAG' column corresponds to twice the number of non-contained contigs.

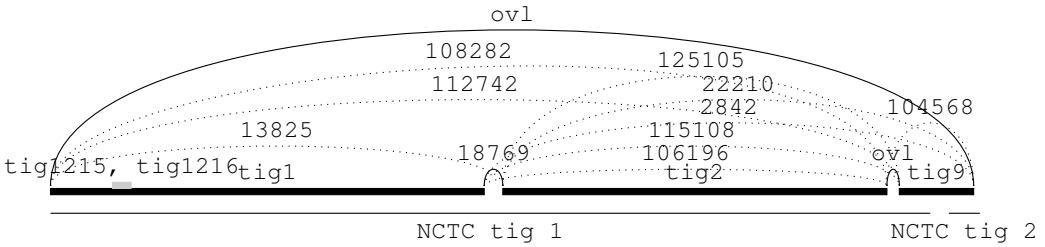

Fig. 3: Mapping of Canu contigs (bold horizontal lines) against NCTC12123 assembly (the two thin horizontal lines). Links between contigs give the length (in bp) of the shortest path in SG between reads at extremities. If two contigs overlap, no length is given and instead the link is labeled with 'ovl'. Plain links are paths that are compatible with the sequential order of contigs given by mapping to the NCTC assembly, and dotted links are all other paths.

Mean number of	
Canu contigs	4.32
Edges in AAG	32.67
Theoretical max. edges in AAG	41.83
Distant edges	28.64
All adjacency edges	4.02
Single adjacency edges	1.16
Multiple adjacency edges	2.86
Dead-ends in Canu contigs	4.94
Dead-ends in AAG, adjacency edges	2.70

Table 2. Average statistics of augmented assembly graphs using a SG built from Minimap2 overlaps on Canu contigs across the 38 NCTC datasets with two or more contigs, after size and classification filters. All rows are as per definitions in Section 4.4. 'Theoretical max. edges': number of possible edges in each AAG. 'Dead-ends in AAG, adjacency edges': number of dead-ends in the AAG when only adjacency edges are considered, i.e. distant edges are deleted.

at this location (see reads mapping in Supplemental Fig. 5) that is likely the cause of assembly fragmentation. Therefore, again in this dataset the path search operation enabled to recover a link between contigs that was discarded by the assembler due to a drop in sequencing coverage.

5.3.2 Path search enables to recover adjacency between contigs
Table 1 reports statistics of paths found between Canu contigs by our method for a subset of 9 NCTC datasets (for the full dataset, see Supplemental Table 3). We first focus on unambiguous contig adjacencies recovered by our pipeline. Single adjacency edges are only found in 6 out of 9 datasets, yet across the entire dataset of 45 samples, 60.4% of all single adjacency edges (43 in total) are found in samples that have a sequencing coverage below 38x, and only 17 single adjacency edges are found in datasets with coverage above 38x. This is likely due to the error-correction step in assemblers that is less effective in low-coverage datasets (even when the true sequencing coverage is given to the assembler as a parameter), which in turn causes assembly fragmentation. Our method therefore enables to recover single adjacency edges between contigs that were fragmented due to this effect.

To measure whether the Canu contig graphs could be used as-is to recover contig order, we counted the number of contig extremities that are not linked to any other extremity (i.e. *dead-ends*). Those are contigs for which no chromosomal order can be reliably inferred. In 35 out of the 45 datasets (7 out of 9 in Table 1), the Canu contig graph has some dead-end extremities (between 1 and 23). In principle dead-ends extremities should not exist in circular bacterial assembly graphs, except for linear chromosomes. Assemblers, here Miniasm and Canu, do not report all true contig adjacencies. In contrast, our method enables to recover some of these adjacencies

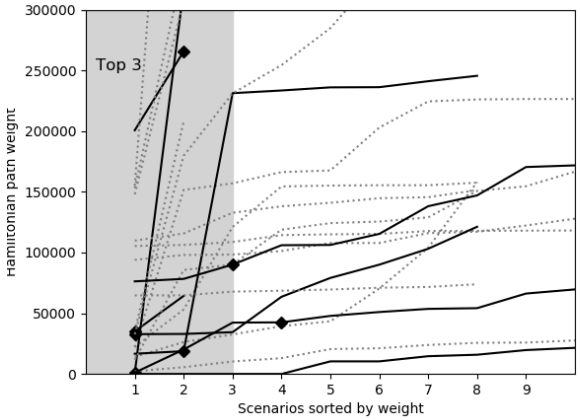

Fig. 4: Weights of scenarios in AAGs. Each curve correspond to the sorted list of Hamiltonian cycles, sorted by weight. If a ground truth is known, a diamond symbol marks the correct assembly scenario. Extended Figure available in Supplementary material 1

and lower the number of dead-ends in 23 out the 37 datasets (and all but one dataset in Table 1).

Table 2 summarizes average AAG statistics over all 38 datasets on `Canu` contigs (per-dataset results in Supplemental Table 3). Results for `Miniasm` contigs are shown in Supplementary Tables 4 and 3. On average, `Canu` contig graphs contain 4.32 nodes (5.86 extremities), among which 4.94 extremities are dead-ends. The AAG enables to reduce the number of dead-end extremities to 2.7 (45% lower), through the discovery of 1.16 single adjacency edges and 2.86 multiple adjacency edges in the AAG per dataset on average. The reduction is also significant for `Miniasm` contigs but not as high (31%, Supp. Table 4). Note that these adjacencies are 'real' in the sense that they are all supported by paths of overlapping reads of total nucleotide length less than 10kbp, yet a number of them may be caused by repetitions. An upper bound on the ability to mine paths in the SG is given by the theoretical maximal number of edges in the AAG (41.83 edges). Our method is on average 78% close to this bound for `Canu` contigs (resp. 90.1% for `Miniasm`) as it discovered 32.67 edges per dataset (resp. 85.1). We note that large fraction (87%) of discovered edges were classified as distant edges, yet the remaining adjacency edges are informative as they significantly contribute to removing dead-ends in the contig graph.

5.3.3 Contig order search retrieves parsimonious assembly scenarios While the work done in the previous section helps to recover contig adjacencies, the presence of multiple adjacency edges due to repetitions often prevents us from unambiguously inferring a contig order. We applied the Hamiltonian cycle procedure presented in Section 4.5 to determine likely contig orderings. Fig. 4 shows orderings sorted by weight across 23 datasets on which the method could successfully be executed (connected AAG, low number of edges).

A ground truth is known in only 8 of those datasets. Among them, the lowest-weight scenario is ranked first in 3 datasets, 2nd in 2 datasets, 3rd in 1, 4th in 1 and 38th in the last one.

These results suggest that the correct assembly scenario is likely to be one of the top predictions made by our parsimonious Hamiltonian cycle procedure. However finding many fragmented datasets that also have a ground truth is inherently difficult, thus further work is needed to confirm this hypothesis. Also, datasets where several scenarios have similar weights (i.e. curves that 'plateau' in Fig 4) will possibly be more challenging to resolve using this method. Yet for many samples with fragmented assemblies, parsimonious assembly scenarios are a promising approach to explore a limited number of hypotheses that could further be validated using long-range PCR to finish the genome.

6 DISCUSSION

We presented a set of concepts to provide novel insights on fragmented long-read bacterial genome assemblies. By searching for paths of overlapping raw reads between extremities of contigs, we construct an *augmented assembly graph* that recovers unreported adjacencies between contigs. We demonstrate several usages of this graph: to provide a more informative representation of fragmented assemblies, to examine repeat structures, and to propose likely contig orderings. In our tests, the AAGs of NCTC datasets recover edges for nearly half (45%) of the dead-end nodes in `Canu` contig graphs, on average. We further show a link between the lowest-weight Hamiltonian cycles in the AAG and the true contig order. We highlight that our method solely relies on the raw data and information produced by assemblers at various stages of their pipelines and, when our contig classification step is skipped, no reference genome nor external information (e.g. genome map, BLAST database) are used.

Our method hinges on directly constructing a string graph on the raw reads, after a relatively conservative chimera removal step. Doing so avoid biases that may be introduced in the read trimming and error-correction steps of an assembler. Indeed, overlaps between reads may become shorter or even absent after error-correction. For instance on the 45 NCTC datasets that we analyzed, the number of edges in SGs built from `Canu` error-corrected reads is reduced by 41.4% compared to the SGs of raw reads. We have classified edges in the AAG, by considering their underlying nucleotide lengths and whether they contain reads that belong to other contigs. To go further, one could define confidence metrics, e.g. based on local graph structures.

Due to a combination of engineering choices and the inherent difficulty of visualizing large assembly graphs, our software has only been tested on bacterial genomes and is unlikely to readily run on larger genomes. However, the techniques presented here (AAG, path search between contig extremities, weighted Hamiltonian cycles) are not specific to bacterial assembly, and should in principle be applicable to small and large eukaryotes. However more work would be needed e.g. to scale path search to thousands of contigs, refine thresholds (contig filter, adjacency edges), handle inter-chromosomal repeats, and an evaluation of the relevance of Hamiltonian cycles for larger genomes.

We stress that our techniques currently do not aim at detecting misassemblies within contigs. We also did not focus on the difficulty of running multiple assembly programs, but we note that the process has previously been reported to be

challenging (Lariviere *et al.*, 2018). Our work is also orthogonal to assembly reconciliation (Alhakami *et al.*, 2017), which consists of constructing a higher-contiguity assembly by merging the results of multiple assemblers.

No attempt was made to optimize the detection of overlaps between reads though this could be a direction for improvement. Finally, automatic post-assembly improvements based on the AAG would be a natural extension of this work. One could use the AAG to design an oracle that suggests a limited number of (long-range) PCR experiments for resolving individual repeats.

ACKNOWLEDGEMENTS

This work was supported by an Inria doctoral grant and the INCEPTION project (PIA/ANR-16-CONV-0005). The authors are grateful to Samarth Rangavittal, Monika Cechova, Jason Chin, Jason Hill and Christopher W. Wheat for discussions that led to this project, Gautam Kamath for guidance on reproducing NCTC analyses with `HINGE`, Antoine Limasset and anonymous reviewers for helpful comments on the manuscript.

REFERENCES

Alhakami, H., Mirebrahim, H., and Lonardi, S. (2017). A comparative evaluation of genome assembly reconciliation tools. *Genome biology*, **18**(1), 93.

Antipov, D. *et al.* (2015). hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, **32**(7), 1009–1015.

Bankevich, A. *et al.* (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, **19**(5), 455–477.

Bosi, E. *et al.* (2015). MeDuSa: a multi-draft based scaffolder. *Bioinformatics*, **31**(15), 2443–2451.

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

Chin, C.-S. *et al.* (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, **13**(12), 1050–1054.

Clark, S. C. *et al.* (2013). ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, **29**(4), 435–443.

Ghodsi, M., Hill, C. M., Astrovskaya, I., Lin, H., Sommer, D. D., Koren, S., and Pop, M. (2013). De novo likelihood-based measures for comparing genome assemblies. *BMC research notes*, **6**(1), 334.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome biology*, **14**(5), R47.

Hunt, M., Silva, N. D., Otto, T. D., Parkhill, J., Keane, J. A., and Harris, S. R. (2015). Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biology*, **16**(1).

Kamath, G. M. *et al.* (2017). HINGE: long-read assembly achieves optimal repeat resolution. *Genome Research*, **27**(5), 747–756.

Koren, S. and Phillippy, A. M. (2015). One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, **23**, 110–120.

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

Koster, J. and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, **28**(19), 2520–2522.

Kremer, F. S. *et al.* (2017). Approaches for in silico finishing of microbial genome sequences. *Genetics and molecular biology*, **40**(3), 553–576.

Lariviere, D., Mei, H., Freeberg, M., Taylor, J., and Nekrutenko, A. (2018). Understanding trivial challenges of microbial genomics: An assembly example. *bioRxiv*.

Lau, B. *et al.* (2016). LongISLND:in silicosequencing of lengthy and noisy datatypes. *Bioinformatics*, **32**(24), 3829–3832.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H. (2016). `Minimap2` and `Miniasm`: Fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics*, **32**(14), 2103–2110.

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M. W., Chaisson, M., and Pevzner, P. A. (2016). Assembly of long error-prone reads using de Bruijn graphs. *Proceedings of the National Academy of Sciences*, **113**(52), E8396–E8405.

Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, **12**(8), 733–735.

Marijon, P., Chikhi, R., and Varré, J. (2019). Optimizing the early steps of long-read genome assembly. Manuscript in preparation.

Miller, J. R. *et al.* (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, **24**(24), 2818–2824.

Myers, G. (2005). The fragment assembly string graph. *Bioinformatics*, **21**(suppl_2), ii79–ii85.

Olson, N. D. *et al.* (2017). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in Bioinformatics*.

Parks, D. H. *et al.* (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, pages gr–186072.

Phillippy, A. M., Schatz, M. C., and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology*, **9**(3), R55.

Rahman, A. and Pachter, L. (2013). CGAL: computing genome assembly likelihoods. *Genome biology*, **14**(1), R8.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**(19), 3210–3212.

Treangen, T. J., Abraham, A.-L., Touchon, M., and Rocha, E. P. (2009). Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiology Reviews*, **33**(3), 539–571.

Utturkar, S. M. *et al.* (2017). A case study into microbial genome assembly gap sequences and finishing strategies. *Frontiers in microbiology*, **8**, 1272.

Vezzi, F., Narzisi, G., and Mishra, B. (2012). Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. *PloS one*, **7**(12), e52210.

Walker, B. J. *et al.* (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, **9**(11), e112963.

Wick, R. R. *et al.* (2015). Bandage: interactive visualization ofde novogenome assemblies: Fig. 1. *Bioinformatics*, **31**(20), 3350–3352.

Wick, R. R. *et al.* (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, **13**(6), e1005595.

Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. (2016). DBG2olc: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific Reports*, **6**(1).

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, **29**(21), 2669–2677.

Appendix 1 Subset of reads included in the string graph

We have noticed that paths between the extremities of two contigs (say, contig A and contig B) correspond to one of the three following situations. i) A path may only contains reads that do not map to any other contig (other than the extremities of A and B). ii) A path contains reads that map to another contig C, but not all bases of contig C are covered by such reads (typically only a possibly small portion of C is covered). iii) A path contains reads that fully span another contig C. Situation (iii) is expected to happen for any pair of non-adjacent contigs. In situation (ii), it is assumed that such a path exists because of a spurious overlap or a repetition. Situation (i) is an interesting case of putative contig adjacencies that were not reported in the original assembly graph.

To avoid as much as possible paths of type ii) and iii), we have decided to remove reads from the SG, keeping only the reads mapping at the end of the contigs and the reads not mapping inside contigs. Concretely, we map reads against contigs, using `Minimap2` with option `-map-[ont|pb]` for Nanopore and Pacbio data, respectively. We keep only reads that do not map to a contig, or reads that map withing a contig at positions that intersect with the interval $[0; t]$ or $[c_l - t, c_l]$, where c_l is the contig length, $t = \max(c_l \times 0.2, 10^4)$.

Appendix 2 Contig classification

In order to have a better understanding of the contig graph produced by a given assembler, we wish to filter out contigs that are not of chromosomal origin. We compare each contig against the *nr* database using Megablast (Morgulis et al., 2008), and classify a contig as chromosomal if its length is greater than 1 Mb, or is such that 80% of the first 50 Megablast hits map to a complete bacterial genome. We use the same second criterion to classify whether a contig is of plasmid origin, regardless of its size. Remaining unclassified contigs are classified as of undefined origin. In addition, we flag as *containment* contigs those which map (using `Minimap2`) over at least 75% of their length to another contig.

Appendix 3 On whether `Canu` contig fragmentation can be solved using `Miniasm` contigs

To check if `Miniasm` contigs could possibly enable to order and fill gaps between `Canu` contigs, we performed an assembly using the `Minimap2` and `Miniasm` pipeline using both the `Canu` contigs and the `Miniasm` contigs as input (to be clear: no reads were used as input to this assembly, only two contig sets). To allow `Minimap2` to find shorter matches, mapping of `Miniasm` contigs against `Canu` contigs was performed with the following parameters: `-x map-pb -m 25 -n 2`. To avoid `Miniasm` filtering overlaps, we ran it with the following parameters: `-1 -2 -s 1000 -c 0`.

We ran this pipeline on all datasets, and counted the number of times that a `Miniasm` contig overlaps with two `Canu` contigs. We also counted the number of contigs generated by `Miniasm` using the overlap created at the previous step. Results are summarized in Supplementary Table 1.

NCTC ID	number of genomic contig `Canu`	`Miniasm`	number of `Miniasm` contigs that overlap two `Canu` contigs	number of merged contigs contigs from `Miniasm`/`Canu` overlaps
NCTC10006	3	7	0	0
NCTC10332	12	22	1	0
NCTC10444	7	5	0	0
NCTC10702	3	2	0	0
NCTC10766	13	7	1	0
NCTC10794	7	5	0	0
NCTC10988	10	9	0	0
NCTC11126	7	15	5	0
NCTC11343	12	10	2	1
NCTC11360	26	25	1	0
NCTC11435	8	6	2	0
NCTC11800	7	3	1	0
NCTC11872	7	13	3	0
NCTC12123	5	3	2	0
NCTC12126	13	15	4	0
NCTC12131	16	77	3	0
NCTC12132	2	4	1	0
NCTC12146	3	1	0	0
NCTC12694	21	123	1	0
NCTC12841	16	1	0	0
NCTC12993	5	2	0	0
NCTC12998	3	4	0	0
NCTC13095	3	2	0	0
NCTC13125	6	7	0	0
NCTC13348	25	17	3	1
NCTC13463	5	4	1	0
NCTC13543	3	3	0	0
NCTC4672	68	16	5	1
NCTC5050	4	4	0	0
NCTC5053	8	11	2	1
NCTC5055	143	20	0	0
NCTC7922	13	9	4	1
NCTC8179	15	15	1	0
NCTC8500	3	1	0	1
NCTC8684	5	2	0	0
NCTC9075	7	3	2	0
NCTC9078	4	2	0	0
NCTC9098	8	6	3	0
NCTC9111	9	13	0	0
NCTC9112	7	15	10	0
NCTC9184	141	17	0	0
NCTC9645	31	76	9	3
NCTC9646	8	9	3	1
NCTC9695	2	1	0	0

Table 1: The pipeline described section Appendix 3 found more than one overlap between `Canu` contigs with `Miniasm` contig for 24 over 45 datasets. When these overlaps are re-assembled using `Miniasm`, one or more merged contigs are produced in only 8 out of 45 datasets.

Appendix 4 Assembly summary

Tables 2 and 3 report our complete results for the 45 NCTC datasets.

NCTC ID	species	cov	NCTC contigs			HINGE	Canu contigs			Miniasm contigs		
			chr	pld	und	status	chr	pld	und	chr	pld	und
NCTC10006	*E. aerogenes*	56	1	0	0	MAF	3	0	0	7	0	2
NCTC10332	*P. aeruginosa*	36	1	0	0	MAF	12	0	0	22	0	1
NCTC10444	*E. coli*	61	1	0	0	MAF	7	0	0	5	0	1
NCTC10702	*S. aureus*	24	1	1	1	MAF	3	3	0	2	1	2
NCTC10766	*E. alkalescens*	37	0	0	11	MA	13	7	3	7	2	5
NCTC10794	*H. parahaemolyticus*	26	0	0	3	MAF	7	0	2	5	0	2
NCTC10988	*S. aureus*	87	0	0	13	MA	10	0	26	9	0	4
NCTC11126	*E. coli*	50	2	0	0	FALC	7	0	2	15	0	18
NCTC11343	*S. multivorum*	22	0	0	11	MAF	12	0	0	10	0	0
NCTC11360	*S. agalactiae*	3	0	0	3	MAF	26	0	17	25	0	1
NCTC11435	*V. mimicus*	60	0	0	3	MA*	8	0	0	6	0	2
NCTC11800	*P. stuartii*	32	0	0	4	MA	7	0	0	3	0	0
NCTC11872	*H. influenzae*	27	0	0	11	MAF	7	0	3	13	0	1
NCTC12123	*E. asburiae*	64	2	3	0	FAMT	5	4	1	3	1	1
NCTC12126	*E.rcancerogenus*	42	6	1	0	MAF	13	1	4	15	0	10
NCTC12131	*Y. regensburgei*	41	3	0	0	MAF	16	0	0	77	0	0
NCTC12132	*M. wisconsensis*	86	1	0	0	MAF	2	0	2	4	0	0
NCTC12146	*Klebsiella terrigena*	11	0	0	2	MAF	3	0	1	1	0	2
NCTC12694	*S. enterica*	19	0	0	121	MAF	21	3	0	123	2	0
NCTC12841	*S. pyogenes*	75	*	*	*	MA	16	0	0	1	0	2
NCTC12993	*K. cryocrescens*	46	5	1	0	FCA	5	4	0	2	3	0
NCTC12998	*R. planticola*	41	1	1	0	MAF	3	2	4	4	1	2
NCTC13095	*K. planticola*	38	1	1	3	FCA	3	3	0	2	0	1
NCTC13125	*E. coli*	49	1	2	4	MAF	6	3	1	7	2	1
NCTC13348	*S. enterica*	41	4	2	0	MAF	25	1	1	17	1	2
NCTC13463	*E. coli*	62	1	1	4	MAF	5	2	2	4	1	3
NCTC13543	*R. radiobacter*	31	0	0	12	MA	3	2	2	3	3	0
NCTC4672	*S. uberis*	10	0	0	3	MAF	68	0	8	16	0	1
NCTC5050	*K. pneumoniae*	54	2	3	0	MAF	4	2	3	4	3	2
NCTC5053	*K. pneumoniae*	28	0	0	7	MAF	8	5	1	11	5	1
NCTC5055	*K. pneumoniae*	69	1	0	2	MAF	143	8	3	20	3	1
NCTC7152	*E. coli*	49	1	0	4	MAF	2	3	5	1	1	3
NCTC7922	*E. coli*	26	0	0	6	MAF	13	3	4	9	2	1
NCTC8179	*E. coli*	36	1	1	3	MAF	15	4	4	15	2	0
NCTC8500	*E. coli*	29	1	1	0	MAF	3	1	1	1	1	5
NCTC8684	*C. violaceum*	36	0	0	3	MAF	5	0	0	2	0	1
NCTC9075	*E. coli*	35	1	0	3	MAF	7	0	14	3	0	1
NCTC9078	*E. coli*	55	1	2	2	MA	4	3	1	2	1	2
NCTC9098	*E. coli*	56	1	1	2	MAF	8	0	1	6	2	2
NCTC9111	*E. coli*	62	1	1	9	MAF	9	6	2	13	3	1
NCTC9112	*E. coli*	69	1	0	0	MAF	7	0	5	15	0	1
NCTC9184	*Klebsiella sp.*	6	0	0	179	MAF	141	5	0	17	0	4
NCTC9645	*K. pneumoniae*	17	0	0	16	MAF	31	10	1	76	4	2
NCTC9646	*K. aerogenes*	24	*	*	*	MAF	8	3	3	9	5	1
NCTC9695	*C. violaceum*	34	1	0	0	MAF	2	9	3	1	0	1
	T. roseus	20	*	*	*	*	3	0	0	6	0	0

Table 2: Datasets from the NCTC project chosen for analysis (the last row corresponds to our simulated dataset). For each sample, the coverage (cov) is given as well as the number of contigs and their assignment; chr: number of chromosomal contigs, pld: number of plasmid contigs, und: number of other contigs. For two datasets (NCTC12841 and NCTC9646) the NCTC project does not yet provide an assembly ("Pending"). For `Canu` and `Miniasm`, a classification similar to the one of NCTC is given (see text). We reported `HINGE` classification; FALC: Finished assembly (lacking circularization), FA: Finished assembly, MA: Mis-assembly, MA*: labeled as misassembled but actually correctly solved as 2 chromosomes, FCA: Finished circular assembly, MAF: Mis-assembly/Fragmented, FAMT: Finished assembly with multiple traversals.

NCTC ID	Canu		dead-ends with adj. edge	Edges in the AAG					
	contigs	dead-ends		total AAG	theoretical max. edges	distant edges	adjacency edges total	single	multiple
NCTC10006	2	2	2	4	4	2	2	2	0
NCTC10332	4	8	4	24	24	21	3	0	3
NCTC10444	4	3	3	24	24	18	6	0	6
NCTC10702	2	4	0	4	4	4	0	0	0
NCTC10766	4	6	2	24	24	22	2	2	0
NCTC10794	3	5	0	12	12	12	0	0	0
NCTC10988	1	0	0	0	0	0	0	0	0
NCTC11126	4	4	3	20	24	15	5	0	5
NCTC11343	7	6	3	72	84	66	6	1	5
NCTC11360	3	6	0	12	12	12	0	0	0
NCTC11435	5	4	4	40	40	35	5	2	3
NCTC11800	2	0	0	3	4	1	2	2	0
NCTC11872	5	6	4	40	40	36	4	4	0
NCTC12123	3	4	3	12	12	9	3	1	2
NCTC12126	6	7	6	36	60	26	10	0	10
NCTC12131	8	6	6	83	112	60	23	0	23
NCTC12132	2	4	2	4	4	3	1	1	0
NCTC12146	2	4	0	4	4	4	0	0	0
NCTC12694	10	20	6	61	180	58	3	3	0
NCTC12841	1	0	0	0	0	0	0	0	0
NCTC12993	2	4	2	4	4	3	1	1	0
NCTC12998	1	2	0	0	0	0	0	0	0
NCTC13095	2	2	0	4	4	3	1	1	0
NCTC13125	3	0	0	12	12	8	4	0	4
NCTC13348	7	7	0	75	84	68	7	0	7
NCTC13463	2	0	0	3	4	1	2	2	0
NCTC13543	2	2	0	4	4	4	0	0	0
NCTC4672	5	8	4	32	40	28	4	0	4
NCTC5050	3	6	6	12	12	9	3	3	0
NCTC5053	5	6	2	32	40	28	4	1	3
NCTC5055	1	2	0	0	0	0	0	0	0
NCTC7152	1	0	0	0	0	0	0	0	0
NCTC7922	6	3	2	60	60	56	4	2	2
NCTC8179	7	4	0	84	84	79	5	3	2
NCTC8500	1	2	0	0	0	0	0	0	0
NCTC8684	1	2	0	0	0	0	0	0	0
NCTC9075	6	8	7	60	60	54	6	4	2
NCTC9078	2	0	0	2	4	1	1	1	0
NCTC9098	4	1	1	24	24	16	8	0	8
NCTC9111	3	2	2	12	12	8	4	0	4
NCTC9112	4	0	0	24	24	14	10	0	10
NCTC9184	0	0	0	0	0	0	0	0	0
NCTC9645	14	23	5	244	364	238	6	3	3
NCTC9646	5	8	4	40	40	37	3	3	0
NCTC9695	2	0	0	2	4	1	1	1	0
Summary	3.71	4.24	1.84	26.86	34.4	23.55	3.31	0.95	2.35

Table 3: Assembly graph statistics for a selection of 45 fragmented assemblies from the NCTC project. `Canu` assembly graph statistics: number of contigs, number of dead-end extremities. AAG statistics: theoretical maximal number number of edges. Note that for some of the most fragmented datasets (e.g. NCTC9184), none of the contigs pass the 100 Kbp length threshold, hence the AAG is empty.

Mean number of	
`Miniasm` contigs	5.8
Edges in AAG	85.1
Theoretical max. edges in AAG	94.4
Distant edges	83.12
All adjacency edges	1.98
Single adjacency edges	1.51
Multiple adjacency edges	0.46
Dead-ends in `Miniasm` contigs	11.61
Dead-ends in AAG, adjacency edges	7.95

Table 4: Average statistics of augmented assembly graphs using a SG built from `Minimap2` overlaps on `Miniasm` contigs across the 37 NCTC datasets with two or more contigs, after size and classification filters. All rows are as per definitions in Section 4.4. 'Theoretical max. edges': number of possible edges in each AAG. 'Dead-ends in AAG, adjacency edges': number of dead-ends in the AAG when only adjacency edges are considered, i.e. distant edges are deleted.

Appendix 5

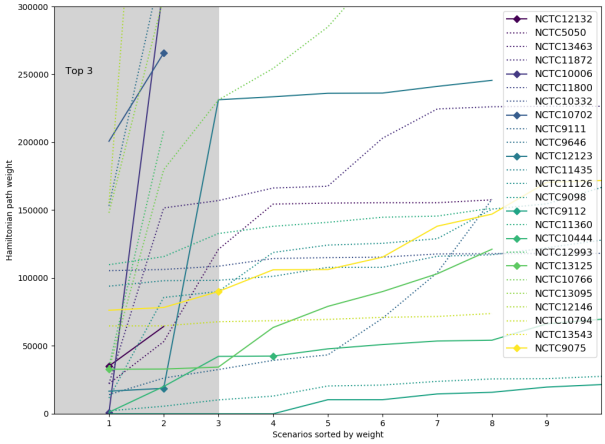

Figure 1: Weights of scenarios in AAGs. Each curve correspond to the sorted list of Hamiltonian cycles, sorted by weight. If a ground truth is known, a diamond symbol marks the correct assembly scenario

NCTC ID	Miniasm		dead-ends with adj. edge	Edges in the AAG					
	contigs	dead-ends		total AAG	theoretical max. edges	distant edges	adjacency edges total	single	multiple
NCTC10006	7	14	11	84	84	78	6	4	2
NCTC10332	14	28	12	364	364	358	6	6	0
NCTC10444	5	10	2	40	40	39	1	1	0
NCTC10702	1	2	0	0	0	0	0	0	0
NCTC10766	5	10	8	40	40	35	5	2	3
NCTC10794	3	6	0	12	12	12	0	0	0
NCTC10988	3	6	2	12	12	11	1	1	0
NCTC11126	11	22	8	200	220	196	4	4	0
NCTC11343	3	6	0	8	12	8	0	0	0
NCTC11360	7	14	2	84	84	83	1	1	0
NCTC11435	6	12	4	33	60	31	2	2	0
NCTC11800	3	6	2	8	12	7	1	1	0
NCTC11872	9	18	10	113	144	108	5	5	0
NCTC12123	2	4	4	4	4	1	3	0	3
NCTC12126	11	22	10	146	220	141	5	5	0
NCTC12131	17	34	2	512	544	511	1	1	0
NCTC12132	2	4	0	4	4	4	0	0	0
NCTC12146	1	2	0	0	0	0	0	0	0
NCTC12694	0	0	0	0	0	0	0	0	0
NCTC12841	1	2	0	0	0	0	0	0	0
NCTC12993	2	4	2	4	4	3	1	1	0
NCTC12998	4	8	2	24	24	23	1	1	0
NCTC13095	1	2	0	0	0	0	0	0	0
NCTC13125	5	10	4	32	40	30	2	2	0
NCTC13348	15	30	19	420	420	410	10	8	2
NCTC13463	4	8	2	18	24	17	1	1	0
NCTC13543	2	4	0	4	4	4	0	0	0
NCTC4672	8	16	*	*	*	*	*	*	*
NCTC5050	4	8	5	24	24	20	4	0	4
NCTC5053	9	18	8	113	144	108	5	2	3
NCTC5055	1	2	0	0	0	0	0	0	0
NCTC7152	2	4	0	4	4	4	0	0	0
NCTC7922	6	12	2	50	60	49	1	1	0
NCTC8179	11	22	4	220	220	218	2	2	0
NCTC8500	1	2	0	0	0	0	0	0	0
NCTC8684	2	4	0	2	4	2	0	0	0
NCTC9075	2	4	2	4	4	3	1	1	0
NCTC9078	2	4	0	4	4	4	0	0	0
NCTC9098	5	10	3	32	40	30	2	0	2
NCTC9111	11	22	2	220	220	219	1	1	0
NCTC9112	10	20	4	180	180	178	2	2	0
NCTC9184	0	0	0	0	0	0	0	0	0
NCTC9645	16	32	2	366	480	365	1	1	0
NCTC9646	7	14	6	72	84	69	3	3	0
NCTC9695	1	2	0	0	0	0	0	0	0
Summary	5.32	10.6	3.27	78.6	87.3	76.8	1.77	1.34	0.432

Table 5: Assembly graph statistics for a selection of 45 fragmented assemblies from the NCTC project. `Miniasm` assembly graph statistics: number of contigs, number of dead-end extremities. AAG statistics: theoretical maximal number number of edges. Note that for some of the most fragmented datasets (e.g. NCTC9184), none of the contigs pass the 100 Kbp length threshold, hence the AAG is empty. '*' denotes dataset for which the result is not available.

Appendix 6 Contigs length and classification

Canu

Dataset	Contig name	Classification	Length
NCTC10006	tig00000055	chromosomal	635691
	tig00001802	chromosomal	4649423
	tig00001803	chromosomal	11996
	total chromosomal length		5297110
NCTC10332	tig00000001	chromosomal	3474338
	tig00000002	chromosomal	30165
	tig00000049	chromosomal	477163
	tig00000076	chromosomal	781581
	tig00000121	chromosomal	2609
	tig00000123	chromosomal	2461
	tig00000125	chromosomal	2452
	tig00009835	chromosomal	2395
	tig00009836	chromosomal	1564849
	tig00009837	chromosomal	11088
	tig00009838	chromosomal	2340
	tig00009839	chromosomal	2302
	total chromosomal length		6353743
NCTC10444	tig00000085	chromosomal	16691
	tig00000105	chromosomal	989155
	tig00000671	chromosomal	2391267
	tig00000672	chromosomal	14372
	tig00000673	chromosomal	1333749
	tig00000674	chromosomal	9774
	tig00000675	chromosomal	603044
	total chromosomal length		5358052
NCTC10702	tig00000001	chromosomal	1882575
	tig00000002	chromosomal	1048854
	tig00000080	chromosomal	70302
	total chromosomal length		3001731
	tig00000200	plasmidic	49994
	tig00001328	plasmidic	28893
	tig00001329	plasmidic	7575
	tig00000084	undefined	30012
	tig00000087	undefined	2442
	tig00000199	undefined	20259
NCTC10766	tig00000009	chromosomal	35058
	tig00000021	chromosomal	331047
	tig00001907	chromosomal	4740
	tig00001908	chromosomal	3602512
	tig00001909	chromosomal	15279
	tig00001910	chromosomal	700851
	tig00001911	chromosomal	14965
	tig00001912	chromosomal	10378
	tig00001913	chromosomal	20674
	tig00001915	chromosomal	10586
	tig00001916	chromosomal	9467
	tig00001921	chromosomal	7453
	tig00001922	chromosomal	710378
	total chromosomal length		5473388
	tig00000032	plasmidic	91068
	tig00000038	plasmidic	6441
	tig00000057	plasmidic	49757
	tig00001917	plasmidic	30098
	tig00001918	plasmidic	12494
	tig00001919	plasmidic	8557
	tig00001920	plasmidic	22116
	tig00000035	undefined	7262
	tig00000036	undefined	3368
NCTC10794	tig00000006	chromosomal	54912
	tig00000027	chromosomal	3322
	tig00000081	chromosomal	92448
	tig00000108	chromosomal	3328
	tig00000189	chromosomal	480759
	tig00000190	chromosomal	290320
	tig00004951	chromosomal	591102
	total chromosomal length		1516191
	tig00000003	undefined	186664
	tig00000014	undefined	105799
	tig00000039	undefined	180586
	tig00000040	undefined	18002
	tig00000042	undefined	42405
	tig00000047	undefined	41214
	tig00000067	undefined	3608
	tig00000072	undefined	3493
	tig00000098	undefined	3650
	tig00000100	undefined	3534
	tig00000102	undefined	3498
	tig00000110	undefined	1327
	tig00000114	undefined	3322
	tig00000116	undefined	3800
	tig00000187	undefined	16904
	tig00000188	undefined	12596
	tig00000191	undefined	12836
	tig00000192	undefined	12673
	tig00004950	undefined	4641
	tig00004952	undefined	105999
	tig00004953	undefined	11722
	tig00004954	undefined	6412
	tig00004955	undefined	9623
	tig00000096	none	5730
	tig00000112	none	1434
	tig00000186	none	5922
NCTC10988	tig00000006	chromosomal	25636
	tig00000279	chromosomal	3040963
	tig00000896	chromosomal	22144
	tig00000897	chromosomal	19058
	tig00000898	chromosomal	14604
	tig00000899	chromosomal	14371
	tig00000900	chromosomal	13453
	tig00000901	chromosomal	19223
	tig00000902	chromosomal	14801
	tig00000903	chromosomal	17641
	total chromosomal length		3201894
	tig00000088	undefined	4456
	tig00000105	undefined	36448
NCTC11126	tig00000037	chromosomal	577906
	tig00000074	chromosomal	666697
	tig00000192	chromosomal	1514971
	tig00000193	chromosomal	6302
	tig00000194	chromosomal	2066502
	tig00003788	chromosomal	8944
	tig00003789	chromosomal	46666
	total chromosomal length		4887988
NCTC11343	tig00000004	chromosomal	11272
	tig00000067	chromosomal	117209
	tig00000083	chromosomal	258828
	tig00000095	chromosomal	5614
	tig00000272	chromosomal	249754
	tig00000291	chromosomal	226277
	tig00000726	chromosomal	2208641
	tig00005693	chromosomal	12001
	tig00005694	chromosomal	876928
	tig00005696	chromosomal	3878
	tig00005698	chromosomal	8934
	tig00005699	chromosomal	214339
	total chromosomal length		4193675
	tig00000047	undefined	470559
	tig00000158	undefined	3400
	tig00000261	undefined	3105
	tig00000357	undefined	55998
	tig00000360	undefined	46002
	tig00000381	undefined	1809
	tig00000727	undefined	136723
	tig00000728	undefined	6803
	tig00000729	undefined	19083

	tig00000730	undefined	6003
	tig00005695	undefined	783131
	tig00005697	undefined	20459
	tig00005700	undefined	12630
	tig00005701	undefined	4319
	tig00005702	undefined	4350
	tig00005703	undefined	325592
	tig00005704	undefined	8727
NCTC11360	tig00000001	chromosomal	856759
	tig00000002	chromosomal	167905
	tig00000023	chromosomal	2941
	tig00000024	chromosomal	74373
	tig00000036	chromosomal	2775
	tig00000039	chromosomal	4498
	tig00000040	chromosomal	90923
	tig00000044	chromosomal	93819
	tig00000059	chromosomal	3339
	tig00000061	chromosomal	5726
	tig00000067	chromosomal	69601
	tig00000084	chromosomal	2711
	tig00000115	chromosomal	85300
	tig00000116	chromosomal	6833
	tig00000117	chromosomal	89165
	tig00000118	chromosomal	7948
	tig00000119	chromosomal	73822
	tig00000121	chromosomal	61437
	tig00000122	chromosomal	11175
	tig00000123	chromosomal	7718
	tig00002040	chromosomal	3708
	tig00002041	chromosomal	446407
	tig00002042	chromosomal	2667
	tig00002043	chromosomal	3997
	tig00002044	chromosomal	6676
	tig00002045	chromosomal	6999
	total chromosomal length		2189222
NCTC11435	tig00000001	chromosomal	1450647
	tig00000002	chromosomal	1627001
	tig00000267	chromosomal	3308
	tig00001171	chromosomal	213960
	tig00001172	chromosomal	11255
	tig00001173	chromosomal	260680
	tig00001174	chromosomal	13563
	tig00001175	chromosomal	871963
	total chromosomal length		4452377
NCTC11800	tig00000003	chromosomal	2775
	tig00000100	chromosomal	2617
	tig00000108	chromosomal	2163
	tig00000110	chromosomal	2403
	tig00000228	chromosomal	797974
	tig00000229	chromosomal	7782
	tig00003669	chromosomal	3650645
	total chromosomal length		4466359
	tig00000002	undefined	2722
	tig00000104	undefined	2273
	tig00003670	undefined	10818
NCTC11872	tig00000016	chromosomal	486488
	tig00000035	chromosomal	414114
	tig00000200	chromosomal	305814
	tig00000201	chromosomal	6170
	tig00000202	chromosomal	554139
	tig00000203	chromosomal	6679
	tig00000204	chromosomal	106287
	total chromosomal length		1879691
	tig00000072	undefined	1688
NCTC12123	tig00000001	chromosomal	2025792
	tig00000002	chromosomal	2402021
	tig00000009	chromosomal	319720
	tig00001215	chromosomal	17954
	tig00001216	chromosomal	19022
	total chromosomal length		4784509
	tig00000045	plasmidic	7248
	tig00001219	plasmidic	14495
	tig00001220	plasmidic	11169
	tig00001221	plasmidic	12838
	tig00000003	undefined	7552
	tig00000036	undefined	2048
	tig00001217	undefined	44732
	tig00001218	undefined	4652
NCTC12126	tig00000002	chromosomal	2504
	tig00000003	chromosomal	6347
	tig00000005	chromosomal	312284
	tig00000018	chromosomal	697355
	tig00000041	chromosomal	180413
	tig00000088	chromosomal	980155
	tig00000103	chromosomal	2869
	tig00000144	chromosomal	58545
	tig00000151	chromosomal	9045
	tig00000154	chromosomal	4231
	tig00000255	chromosomal	1991519
	tig00000256	chromosomal	3006
	tig00000257	chromosomal	620710
	total chromosomal length		4868983
	tig00000119	plasmidic	168880
NCTC12131	tig00000004	chromosomal	585202
	tig00000022	chromosomal	2720
	tig00000052	chromosomal	782381
	tig00000129	chromosomal	3215
	tig00000133	chromosomal	2760
	tig00000260	chromosomal	277244
	tig00000261	chromosomal	6991
	tig00000262	chromosomal	654135
	tig00000263	chromosomal	7044
	tig00000264	chromosomal	44446
	tig00000265	chromosomal	6070
	tig00000266	chromosomal	658311
	tig00000267	chromosomal	173281
	tig00000268	chromosomal	6652
	tig00000269	chromosomal	735339
	tig00000271	chromosomal	839355
	total chromosomal length		4785146
	tig00000272	undefined	11590
	tig00000273	undefined	3170
NCTC12132	tig00000001	chromosomal	2583454
	tig00000002	chromosomal	756442
	total chromosomal length		3339896
	tig00000004	undefined	20873
NCTC12146	tig00000001	chromosomal	4385596
	tig00001748	chromosomal	15170
	tig00001749	chromosomal	1248170
	total chromosomal length		5648936
NCTC12694	tig00000001	chromosomal	1305929
	tig00000004	chromosomal	723799
	tig00000010	chromosomal	270213
	tig00000013	chromosomal	244711
	tig00000015	chromosomal	205059
	tig00000017	chromosomal	163002
	tig00000019	chromosomal	200318
	tig00000021	chromosomal	138348
	tig00000028	chromosomal	101438
	tig00000031	chromosomal	87449
	tig00000032	chromosomal	63734
	tig00000035	chromosomal	90673
	tig00000038	chromosomal	41457
	tig00000040	chromosomal	64898
	tig00000042	chromosomal	69114
	tig00000045	chromosomal	37727
	tig00000047	chromosomal	27321
	tig00000052	chromosomal	2331
	tig00000091	chromosomal	753465
	tig00000092	chromosomal	2291
	tig00000093	chromosomal	2280
	total chromosomal length		4595557

	tig00000050	plasmidic	1930
	tig00006898	plasmidic	5801
	tig00006899	plasmidic	63765
NCTC12841	tig00000004	chromosomal	12036
	tig00000005	chromosomal	1851
	tig00000007	chromosomal	2368
	tig00000047	chromosomal	1630
	tig00000050	chromosomal	1416
	tig00000052	chromosomal	2797
	tig00000054	chromosomal	2185
	tig00000058	chromosomal	1348
	tig00000060	chromosomal	1588
	tig00000066	chromosomal	2323
	tig00000257	chromosomal	1926784
	tig00000258	chromosomal	11427
	tig00032866	chromosomal	17087
	tig00032867	chromosomal	11198
	tig00032868	chromosomal	1405
	tig00032869	chromosomal	1416
	total chromosomal length		1998859
NCTC12993	tig00000002	chromosomal	2655515
	tig00002251	chromosomal	2377976
	tig00002252	chromosomal	8006
	tig00002253	chromosomal	9235
	tig00002254	chromosomal	11903
	total chromosomal length		5062635
	tig00000055	plasmidic	12328
	tig00000063	plasmidic	5676
	tig00000064	plasmidic	2730
	tig00000113	plasmidic	5891
	tig00000052	undefined	222246
	tig00000114	undefined	4385
	tig00002255	undefined	9923
	tig00002256	undefined	13795
NCTC12998	tig00000002	chromosomal	2569
	tig00002880	chromosomal	5608109
	tig00002881	chromosomal	9135
	total chromosomal length		5619813
	tig00002882	plasmidic	126740
	tig00002883	plasmidic	7454
NCTC13095	tig00000036	chromosomal	2168596
	tig00000037	chromosomal	8008
	tig00000038	chromosomal	3511453
	total chromosomal length		5688057
	tig00000015	plasmidic	166342
	tig00001684	plasmidic	124320
	tig00001685	plasmidic	18225
	tig00000003	none	21738
NCTC13125	tig00000001	chromosomal	4777685
	tig00000003	chromosomal	461931
	tig00000408	chromosomal	263450
	tig00000409	chromosomal	19433
	tig00001778	chromosomal	18427
	tig00001779	chromosomal	24134
	total chromosomal length		5565060
	tig00000080	plasmidic	105599
	tig00000081	plasmidic	120877
	tig00000083	plasmidic	18752
	tig00000084	undefined	56975
	tig00000099	undefined	1213
NCTC13348	tig00000012	chromosomal	2875
	tig00000029	chromosomal	163558
	tig00000045	chromosomal	2613
	tig00000114	chromosomal	3490
	tig00000124	chromosomal	2641
	tig00000162	chromosomal	87300
	tig00000171	chromosomal	2696
	tig00000186	chromosomal	2783
	tig00000348	chromosomal	742809
	tig00000349	chromosomal	7200
	tig00000350	chromosomal	898431
	tig00000351	chromosomal	4061
	tig00000352	chromosomal	224612
	tig00000353	chromosomal	201081
	tig00000356	chromosomal	2525
	tig00005291	chromosomal	1458800
	tig00005292	chromosomal	3871
	tig00005293	chromosomal	1173550
	tig00005294	chromosomal	6497
	tig00005295	chromosomal	8532
	tig00005296	chromosomal	9588
	tig00005297	chromosomal	3796
	tig00005298	chromosomal	3770
	tig00005299	chromosomal	6030
	tig00005300	chromosomal	4132
	total chromosomal length		5027241
	tig00000183	plasmidic	99046
	tig00000196	undefined	4009
	tig00000355	undefined	2810
NCTC13463	tig00000066	chromosomal	4612761
	tig00000067	chromosomal	15891
	tig00000068	chromosomal	473422
	tig00000070	chromosomal	11585
	tig00000071	chromosomal	9027
	total chromosomal length		5122686
	tig00000024	plasmidic	99437
	tig00000028	plasmidic	3907
	tig00000026	undefined	9287
	tig00000069	undefined	63008
NCTC13543	tig00000001	chromosomal	2912152
	tig00000044	chromosomal	20274
	tig00000092	chromosomal	1100488
	total chromosomal length		4032914
	tig00000037	plasmidic	71251
	tig00000039	plasmidic	27238
	tig00000024	undefined	489748
	tig00000034	undefined	174464
	tig00000042	undefined	31750
	tig00000047	undefined	3595
	tig00000049	undefined	6247
	tig00000057	undefined	3104
	tig00000093	undefined	8383
	tig00000094	undefined	985883
NCTC4672	tig00000005	chromosomal	234563
	tig00000013	chromosomal	183355
	tig00000048	chromosomal	1061
	tig00000049	chromosomal	2728
	tig00000057	chromosomal	1112
	tig00000065	chromosomal	1150
	tig00000084	chromosomal	1399
	tig00000092	chromosomal	3917
	tig00000095	chromosomal	1292
	tig00000124	chromosomal	3541
	tig00000128	chromosomal	1599
	tig00000139	chromosomal	1711
	tig00000144	chromosomal	4980
	tig00000159	chromosomal	3137
	tig00000198	chromosomal	1889
	tig00000233	chromosomal	1995
	tig00000242	chromosomal	7213
	tig00000258	chromosomal	1676
	tig00000262	chromosomal	2808
	tig00000265	chromosomal	1307
	tig00000266	chromosomal	1405
	tig00000269	chromosomal	3265
	tig00000275	chromosomal	1624
	tig00000277	chromosomal	1819
	tig00000278	chromosomal	1609
	tig00000280	chromosomal	1497
	tig00000283	chromosomal	1155
	tig00000288	chromosomal	1859
	tig00000290	chromosomal	1529

	tig00000296	chromosomal	4494
	tig00000297	chromosomal	2018
	tig00000300	chromosomal	1525
	tig00000304	chromosomal	1433
	tig00000306	chromosomal	1315
	tig00000309	chromosomal	1535
	tig00000320	chromosomal	1446
	tig00000323	chromosomal	1479
	tig00000330	chromosomal	1947
	tig00000334	chromosomal	3660
	tig00000338	chromosomal	1749
	tig00000345	chromosomal	1368
	tig00000347	chromosomal	1669
	tig00000349	chromosomal	1420
	tig00000358	chromosomal	1659
	tig00000367	chromosomal	1044
	tig00000380	chromosomal	1237
	tig00000886	chromosomal	59611
	tig00000887	chromosomal	11888
	tig00000888	chromosomal	778559
	tig00000889	chromosomal	130493
	tig00000890	chromosomal	5574
	tig00000891	chromosomal	34698
	tig00000892	chromosomal	1261
	tig00000893	chromosomal	4516
	tig00012913	chromosomal	528826
	tig00012914	chromosomal	3329
	tig00012915	chromosomal	8743
	tig00012916	chromosomal	6640
	tig00012917	chromosomal	8651
	tig00012918	chromosomal	1378
	tig00012919	chromosomal	1701
	tig00012920	chromosomal	1535
	tig00012921	chromosomal	1435
	tig00012922	chromosomal	1444
	tig00012923	chromosomal	1179
	tig00012924	chromosomal	1180
	tig00012925	chromosomal	3460
	tig00012926	chromosomal	3441
	total chromosomal length		2108735
	tig00000046	undefined	3142
	tig00012927	undefined	1024
	tig00012928	undefined	1023
NCTC5050	tig00000001	chromosomal	3626030
	tig00000010	chromosomal	1250471
	tig00000023	chromosomal	227716
	tig00000041	chromosomal	3864
	total chromosomal length		5108081
	tig00000038	plasmidic	82367
	tig00000039	plasmidic	52025
	tig00000037	undefined	117821
NCTC5053	tig00000133	chromosomal	198522
	tig00000255	chromosomal	920215
	tig00000256	chromosomal	5841
	tig00000257	chromosomal	1006535
	tig00000258	chromosomal	6903
	tig00000259	chromosomal	2186965
	tig00003210	chromosomal	6218
	tig00003211	chromosomal	930059
	total chromosomal length		5261258
	tig00000136	plasmidic	112876
	tig00000143	plasmidic	105258
	tig00000146	plasmidic	13447
	tig00000160	plasmidic	9791
	tig00000261	plasmidic	209198
	tig00000260	undefined	9413
	tig00003209	undefined	107411
	tig00003212	undefined	10219
NCTC5055	tig00000055	chromosomal	11815
	tig00000057	chromosomal	12732
	tig00000059	chromosomal	6105
	tig00000060	chromosomal	4943
	tig00000064	chromosomal	5662
	tig00000065	chromosomal	15192
	tig00000070	chromosomal	3156
	tig00000074	chromosomal	4830
	tig00000076	chromosomal	4460
	tig00000077	chromosomal	7113
	tig00000078	chromosomal	5247
	tig00000080	chromosomal	4590
	tig00000081	chromosomal	7472
	tig00000082	chromosomal	2196
	tig00000084	chromosomal	9133
	tig00000094	chromosomal	5354
	tig00000095	chromosomal	3374
	tig00000096	chromosomal	4914
	tig00000097	chromosomal	9470
	tig00000098	chromosomal	6023
	tig00000100	chromosomal	2873
	tig00000101	chromosomal	3992
	tig00000102	chromosomal	3774
	tig00000106	chromosomal	5073
	tig00000107	chromosomal	8708
	tig00000109	chromosomal	6353
	tig00000110	chromosomal	3657
	tig00000112	chromosomal	2278
	tig00000116	chromosomal	3106
	tig00000117	chromosomal	2467
	tig00000119	chromosomal	4337
	tig00000122	chromosomal	3239
	tig00000127	chromosomal	3330
	tig00000129	chromosomal	3810
	tig00000130	chromosomal	8852
	tig00000133	chromosomal	4009
	tig00000151	chromosomal	1816
	tig00000153	chromosomal	4264
	tig00000154	chromosomal	9420
	tig00000155	chromosomal	3231
	tig00000156	chromosomal	3481
	tig00000158	chromosomal	2227
	tig00000159	chromosomal	5958
	tig00000160	chromosomal	3393
	tig00000161	chromosomal	2176
	tig00000162	chromosomal	2694
	tig00000163	chromosomal	2441
	tig00000165	chromosomal	1982
	tig00000167	chromosomal	8049
	tig00000168	chromosomal	3057
	tig00000169	chromosomal	4639
	tig00000171	chromosomal	5174
	tig00000172	chromosomal	4436
	tig00000176	chromosomal	2044
	tig00000177	chromosomal	3065
	tig00000179	chromosomal	5480
	tig00000180	chromosomal	5299
	tig00000181	chromosomal	7740
	tig00000182	chromosomal	3451
	tig00000183	chromosomal	3189
	tig00000184	chromosomal	1334
	tig00000186	chromosomal	3107
	tig00000187	chromosomal	2091
	tig00000189	chromosomal	2580
	tig00000190	chromosomal	1472
	tig00000191	chromosomal	8189
	tig00000192	chromosomal	5362
	tig00000193	chromosomal	3042
	tig00000194	chromosomal	6645
	tig00000195	chromosomal	1695
	tig00000196	chromosomal	1678
	tig00000202	chromosomal	2682
	tig00000203	chromosomal	9552
	tig00000204	chromosomal	3295

tig00000209	chromosomal	6262
tig00000212	chromosomal	6643
tig00000220	chromosomal	3460
tig00000225	chromosomal	2926
tig00000229	chromosomal	5309
tig00000256	chromosomal	3504
tig00000263	chromosomal	1548
tig00000264	chromosomal	2071
tig00000266	chromosomal	7550
tig00000267	chromosomal	1633
tig00000269	chromosomal	2507
tig00000273	chromosomal	3348
tig00000274	chromosomal	3548
tig00000275	chromosomal	4156
tig00000277	chromosomal	3110
tig00000279	chromosomal	4417
tig00000281	chromosomal	3851
tig00000288	chromosomal	5472
tig00000289	chromosomal	4032
tig00000291	chromosomal	3818
tig00000292	chromosomal	4370
tig00000293	chromosomal	3129
tig00000294	chromosomal	2304
tig00000296	chromosomal	3225
tig00000301	chromosomal	7281
tig00000305	chromosomal	8491
tig00000314	chromosomal	5433
tig00000317	chromosomal	3678
tig00000325	chromosomal	1863
tig00000327	chromosomal	3222
tig00000328	chromosomal	5106
tig00000333	chromosomal	3256
tig00000341	chromosomal	1291
tig00000342	chromosomal	2493
tig00000346	chromosomal	1815
tig00000353	chromosomal	2918
tig00000355	chromosomal	4982
tig00000357	chromosomal	2946
tig00000358	chromosomal	1834
tig00000360	chromosomal	2630
tig00000364	chromosomal	3574
tig00000370	chromosomal	2820
tig00000378	chromosomal	8735
tig00000381	chromosomal	3848
tig00000382	chromosomal	2055
tig00000386	chromosomal	2616
tig00000387	chromosomal	1427
tig00000389	chromosomal	1736
tig00000397	chromosomal	5670
tig00000401	chromosomal	2024
tig00000407	chromosomal	3932
tig00000409	chromosomal	4037
tig00000426	chromosomal	1317
tig00000429	chromosomal	5444
tig00000430	chromosomal	3589
tig00001790	chromosomal	22546
tig00001791	chromosomal	4656080
tig00008453	chromosomal	2243
tig00008454	chromosomal	3013
tig00008455	chromosomal	2506
tig00008456	chromosomal	2503
tig00008457	chromosomal	2363
tig00008458	chromosomal	1846
tig00008459	chromosomal	8988
tig00008460	chromosomal	7418
tig00008461	chromosomal	2516
tig00008462	chromosomal	1286
tig00008463	chromosomal	1256
tig00008464	chromosomal	1283
total chromosomal length		5275172
tig00000062	plasmidic	18081
tig00000105	plasmidic	10705
tig00000121	plasmidic	7705
tig00000157	plasmidic	2638
tig00000228	plasmidic	5859
tig00000336	plasmidic	1653
tig00000366	plasmidic	1797
tig00001789	plasmidic	274671
tig00000173	undefined	4757
tig00000270	undefined	5189
tig00000282	undefined	2754
tig00000306	undefined	2137
tig00000308	undefined	5179
NCTC7152	tig00001521	chromosomal	4895392
tig00001522	chromosomal	11663
total chromosomal length		4907055
tig00000020	plasmidic	140100
tig00000021	plasmidic	22029
tig00000023	plasmidic	17571
tig00000004	undefined	12499
tig00001524	undefined	9161
tig00000002	none	14822
tig00001523	none	13250
NCTC7922	tig00000005	chromosomal	30266
tig00000010	chromosomal	231607
tig00000015	chromosomal	8910
tig00000061	chromosomal	624029
tig00000089	chromosomal	224263
tig00000120	chromosomal	118779
tig00000357	chromosomal	3437368
tig00000358	chromosomal	27476
tig00000359	chromosomal	62893
tig00000360	chromosomal	14447
tig00000361	chromosomal	517854
tig00004505	chromosomal	2628
tig00004506	chromosomal	7127
total chromosomal length		5307647
tig00000123	plasmidic	92363
tig00000136	plasmidic	68892
tig00000137	plasmidic	2971
tig00000138	undefined	2688
tig00000140	undefined	8279
tig00000143	undefined	5528
tig00000356	undefined	9292
NCTC8179	tig00000002	chromosomal	32726
tig00000005	chromosomal	34757
tig00000006	chromosomal	156816
tig00000012	chromosomal	932548
tig00000140	chromosomal	32623
tig00000141	chromosomal	1989140
tig00000143	chromosomal	297068
tig00000144	chromosomal	33325
tig00000145	chromosomal	260864
tig00000146	chromosomal	22495
tig00000147	chromosomal	1150072
tig00001520	chromosomal	24836
tig00001521	chromosomal	21995
tig00001522	chromosomal	17732
tig00001523	chromosomal	732378
total chromosomal length		5739375
tig00000063	plasmidic	127915
tig00000065	plasmidic	85310
tig00000066	plasmidic	5132
tig00000069	plasmidic	3833
tig00000142	none	18135
NCTC8500	tig00000001	chromosomal	4654897
tig00000069	chromosomal	14271
tig00000172	chromosomal	2477
total chromosomal length		4671645
tig00000166	plasmidic	61752
NCTC8684	tig00000042	chromosomal	2510
tig00000044	chromosomal	2653

	tig00000096	chromosomal	1675130
	tig00000100	chromosomal	9818
	tig00005015	chromosomal	2002
	total chromosomal length		1692113
	tig00000019	undefined	90777
	tig00000035	undefined	334610
	tig00000040	undefined	2954
	tig00000090	undefined	463211
	tig00000091	undefined	6937
	tig00000092	undefined	226539
	tig00000093	undefined	10565
	tig00000094	undefined	683840
	tig00000095	undefined	8091
	tig00000097	undefined	11829
	tig00000098	undefined	815147
	tig00000099	undefined	582249
	tig00005013	undefined	3845
	tig00005014	undefined	3834
NCTC9075	tig00000001	chromosomal	2771864
	tig00000014	chromosomal	707603
	tig00000055	chromosomal	975632
	tig00000129	chromosomal	250221
	tig00000196	chromosomal	115073
	tig00002929	chromosomal	6892
	tig00002930	chromosomal	441745
	total chromosomal length		5269030
	tig00000200	undefined	67419
NCTC9078	tig00000001	chromosomal	4157901
	tig00000006	chromosomal	11044
	tig00000036	chromosomal	1211
	tig00000051	chromosomal	1033327
	total chromosomal length		5203483
	tig00000025	plasmidic	84831
	tig00000052	plasmidic	15048
	tig00000053	plasmidic	141326
	tig00000050	undefined	13786
NCTC9098	tig00000001	chromosomal	3151410
	tig00000030	chromosomal	19458
	tig00000163	chromosomal	19823
	tig00000526	chromosomal	324234
	tig00000527	chromosomal	19807
	tig00000528	chromosomal	196308
	tig00000529	chromosomal	15991
	tig00000530	chromosomal	1487922
	total chromosomal length		5234953
	tig00000209	none	64136
	tig00000212	none	86222
NCTC9111	tig00000001	chromosomal	4605377
	tig00000032	chromosomal	15239
	tig00000054	chromosomal	151455
	tig00000063	chromosomal	586362
	tig00000064	chromosomal	28263
	tig00000186	chromosomal	5942
	tig00002643	chromosomal	30626
	tig00002644	chromosomal	14812
	tig00002645	chromosomal	16371
	total chromosomal length		5454447
	tig00000120	plasmidic	4002
	tig00000187	plasmidic	88084
	tig00002646	plasmidic	132127
	tig00002648	plasmidic	84308
	tig00002649	plasmidic	16303
	tig00002651	plasmidic	12651
	tig00000118	undefined	3898
	tig00000123	undefined	106160
	tig00002647	undefined	12391
	tig00002650	undefined	10115
	tig00002642	none	17615
NCTC9112	tig00000065	chromosomal	1280329
	tig00000084	chromosomal	1227588
	tig00000705	chromosomal	761814
	tig00000706	chromosomal	23527
	tig00000707	chromosomal	2213829
	tig00001864	chromosomal	23718
	tig00001865	chromosomal	26283
	total chromosomal length		5557088
NCTC9184	tig00000001	chromosomal	44206
	tig00000003	chromosomal	75248
	tig00000005	chromosomal	54514
	tig00000010	chromosomal	38058
	tig00000013	chromosomal	57613
	tig00000015	chromosomal	34569
	tig00000017	chromosomal	29672
	tig00000021	chromosomal	41507
	tig00000022	chromosomal	32208
	tig00000025	chromosomal	33831
	tig00000027	chromosomal	30328
	tig00000028	chromosomal	25544
	tig00000029	chromosomal	31199
	tig00000031	chromosomal	18451
	tig00000032	chromosomal	26706
	tig00000036	chromosomal	27618
	tig00000039	chromosomal	27467
	tig00000040	chromosomal	21778
	tig00000042	chromosomal	28070
	tig00000045	chromosomal	26501
	tig00000046	chromosomal	23919
	tig00000048	chromosomal	35573
	tig00000049	chromosomal	18586
	tig00000051	chromosomal	24356
	tig00000055	chromosomal	35816
	tig00000056	chromosomal	37501
	tig00000057	chromosomal	9675
	tig00000058	chromosomal	25028
	tig00000059	chromosomal	21381
	tig00000060	chromosomal	32086
	tig00000061	chromosomal	22676
	tig00000063	chromosomal	20847
	tig00000065	chromosomal	16377
	tig00000066	chromosomal	22324
	tig00000069	chromosomal	23508
	tig00000071	chromosomal	23542
	tig00000072	chromosomal	24820
	tig00000078	chromosomal	13426
	tig00000082	chromosomal	23417
	tig00000088	chromosomal	17003
	tig00000089	chromosomal	15211
	tig00000090	chromosomal	21564
	tig00000091	chromosomal	10799
	tig00000094	chromosomal	34765
	tig00000095	chromosomal	16175
	tig00000096	chromosomal	28943
	tig00000099	chromosomal	2490
	tig00000102	chromosomal	10959
	tig00000104	chromosomal	15702
	tig00000105	chromosomal	17032
	tig00000113	chromosomal	17463
	tig00000114	chromosomal	24382
	tig00000115	chromosomal	6126
	tig00000116	chromosomal	7311
	tig00000117	chromosomal	6497
	tig00000118	chromosomal	13154
	tig00000119	chromosomal	19876
	tig00000121	chromosomal	17839
	tig00000122	chromosomal	10689
	tig00000124	chromosomal	14467
	tig00000128	chromosomal	16138
	tig00000129	chromosomal	18515
	tig00000134	chromosomal	15758
	tig00000135	chromosomal	7877
	tig00000139	chromosomal	12365
	tig00000141	chromosomal	23830

tig00000145	chromosomal	15645
tig00000147	chromosomal	21070
tig00000148	chromosomal	31094
tig00000158	chromosomal	6592
tig00000159	chromosomal	12026
tig00000160	chromosomal	19542
tig00000161	chromosomal	16653
tig00000162	chromosomal	9525
tig00000163	chromosomal	3503
tig00000164	chromosomal	10038
tig00000168	chromosomal	22095
tig00000171	chromosomal	5815
tig00000172	chromosomal	3557
tig00000173	chromosomal	8034
tig00000174	chromosomal	13049
tig00000175	chromosomal	13166
tig00000176	chromosomal	4913
tig00000177	chromosomal	4186
tig00000182	chromosomal	16661
tig00000184	chromosomal	12911
tig00000187	chromosomal	10310
tig00000191	chromosomal	11302
tig00000193	chromosomal	10014
tig00000199	chromosomal	11611
tig00000201	chromosomal	14360
tig00000204	chromosomal	2382
tig00000210	chromosomal	10068
tig00000212	chromosomal	8977
tig00000223	chromosomal	17229
tig00000240	chromosomal	6878
tig00000241	chromosomal	15069
tig00000242	chromosomal	6620
tig00000245	chromosomal	15723
tig00000246	chromosomal	4700
tig00000248	chromosomal	5491
tig00000250	chromosomal	14542
tig00000252	chromosomal	20309
tig00000253	chromosomal	5109
tig00000254	chromosomal	6407
tig00000255	chromosomal	4126
tig00000263	chromosomal	18307
tig00000264	chromosomal	6065
tig00000269	chromosomal	2756
tig00000272	chromosomal	15386
tig00000273	chromosomal	10403
tig00000276	chromosomal	3194
tig00000280	chromosomal	10412
tig00000289	chromosomal	5925
tig00000297	chromosomal	2750
tig00000301	chromosomal	14266
tig00000305	chromosomal	6556
tig00000311	chromosomal	4992
tig00000315	chromosomal	5174
tig00000316	chromosomal	9510
tig00000318	chromosomal	3586
tig00003367	chromosomal	9616
tig00003368	chromosomal	55674
tig00003369	chromosomal	40990
tig00003370	chromosomal	4425
tig00003371	chromosomal	11626
tig00003372	chromosomal	28757
tig00003373	chromosomal	23908
tig00003374	chromosomal	6632
tig00003375	chromosomal	6460
tig00003376	chromosomal	8901
tig00003377	chromosomal	13617
tig00003378	chromosomal	7801
tig00003379	chromosomal	3818
tig00003380	chromosomal	7550
tig00003381	chromosomal	10743
tig00003382	chromosomal	11567
tig00003383	chromosomal	14003
tig00003384	chromosomal	15516
tig00003385	chromosomal	17588
tig00003386	chromosomal	17512
total chromosomal length		2470164
tig00000107	plasmidic	15044
tig00000166	plasmidic	10162
tig00000186	plasmidic	12869
tig00000188	plasmidic	18137
tig00000299	plasmidic	2027
tig00000180	undefined	13067
NCTC9645	tig00000007	chromosomal	2625
	tig00000011	chromosomal	607255
	tig00000013	chromosomal	599160
	tig00000021	chromosomal	40668
	tig00000024	chromosomal	405420
	tig00000026	chromosomal	317955
	tig00000036	chromosomal	103955
	tig00000037	chromosomal	234258
	tig00000042	chromosomal	220152
	tig00000047	chromosomal	201508
	tig00000052	chromosomal	207208
	tig00000058	chromosomal	2660
	tig00000061	chromosomal	135529
	tig00000094	chromosomal	27162
	tig00000096	chromosomal	18889
	tig00000098	chromosomal	20876
	tig00000101	chromosomal	5995
	tig00000102	chromosomal	1801
	tig00000105	chromosomal	2743
	tig00000109	chromosomal	1646
	tig00000113	chromosomal	1844
	tig00000206	chromosomal	382698
	tig00000207	chromosomal	7336
	tig00000208	chromosomal	1225204
	tig00000209	chromosomal	232251
	tig00000210	chromosomal	16029
	tig00000211	chromosomal	97614
	tig00000219	chromosomal	3443
	tig00000220	chromosomal	80500
	tig00012227	chromosomal	4751
	tig00012228	chromosomal	100031
	total chromosomal length		5309166
	tig00000072	plasmidic	10238
	tig00000086	plasmidic	8968
	tig00000212	plasmidic	82446
	tig00000213	plasmidic	14986
	tig00000214	plasmidic	32714
	tig00000215	plasmidic	99472
	tig00000217	plasmidic	11997
	tig00000218	plasmidic	87460
	tig00000221	plasmidic	2054
	tig00000222	plasmidic	13460
	tig00000035	undefined	168687
	tig00000088	undefined	7153
	tig00000069	none	81493
NCTC9646	tig00000001	chromosomal	3665711
	tig00000002	chromosomal	614927
	tig00000026	chromosomal	206992
	tig00000027	chromosomal	878265
	tig00000047	chromosomal	295064
	tig00000187	chromosomal	2534
	tig00003591	chromosomal	4056
	tig00003592	chromosomal	4764
	total chromosomal length		5672313
	tig00000022	plasmidic	148222
	tig00003589	plasmidic	8057
	tig00003590	plasmidic	6751
	tig00000021	undefined	36388
	tig00000063	undefined	1282
	tig00000065	undefined	1113

NCTC9695	tig00000074	chromosomal	1279605
	tig00000076	chromosomal	1894574
	total chromosomal length		3174179
	tig00000003	undefined	473776
	tig00000004	undefined	204759
	tig00000019	undefined	4937
	tig00000038	undefined	3861
	tig00000040	undefined	2672
	tig00000042	undefined	2170
	tig00000075	undefined	7627
	tig00000077	undefined	7294
	tig00000078	undefined	911580

Table 6: Canu contigs classification per NCTC dataset. Total length of chromosomal contigs is given.

Miniasm

Dataset	Contig name	Classification	Length
NCTC10006	utg000001l	chromosomal	260336
	utg000002l	chromosomal	1081553
	utg000003l	chromosomal	1615186
	utg000004l	chromosomal	1435892
	utg000005l	chromosomal	629371
	utg000006l	chromosomal	301502
	utg000007l	chromosomal	263124
	total chromosomal length		5586964
NCTC10332	utg000002l	chromosomal	213696
	utg000003l	chromosomal	598526
	utg000004l	chromosomal	220355
	utg000005l	chromosomal	687999
	utg000006l	chromosomal	92810
	utg000007l	chromosomal	274321
	utg000008l	chromosomal	889152
	utg000009l	chromosomal	236367
	utg000010l	chromosomal	62450
	utg000011l	chromosomal	436257
	utg000012l	chromosomal	720033
	utg000013l	chromosomal	191547
	utg000014l	chromosomal	301467
	utg000015l	chromosomal	41317
	utg000016l	chromosomal	350649
	utg000017l	chromosomal	273059
	utg000018l	chromosomal	339294
	utg000019l	chromosomal	65630
	utg000020l	chromosomal	81777
	utg000022l	chromosomal	43390
	utg000023l	chromosomal	42183
	utg000024l	chromosomal	16950
	total chromosomal length		6179229
	utg000011l	none	274988
	utg000021l	none	57736
NCTC10444	utg000001l	chromosomal	2018895
	utg000002l	chromosomal	1852358
	utg000003l	chromosomal	240957
	utg000004l	chromosomal	1224505
	utg000005c	chromosomal	234694
	total chromosomal length		5571409
	utg000006c	none	4134
NCTC10702	utg000001c	chromosomal	3036414
	utg000004l	chromosomal	5937
	total chromosomal length		3042351
	utg000003l	plasmidic	36874
	utg000002c	undefined	34724
NCTC10766	utg000001l	chromosomal	424892
	utg000002l	chromosomal	360757
	utg000003l	chromosomal	3136691
	utg000004l	chromosomal	991822
	utg000005l	chromosomal	814775
	utg000007l	chromosomal	17423
	utg000010l	chromosomal	76701
	total chromosomal length		5823061
	utg000006c	plasmidic	56779
	utg000008c	plasmidic	88390
	utg000009l	undefined	7051
	utg000011c	undefined	6079
NCTC10794	utg000001l	chromosomal	686754
	utg000003l	chromosomal	73314
	utg000004l	chromosomal	198317
	utg000005l	chromosomal	344693
	utg000008l	chromosomal	84572
	total chromosomal length		1387650
	utg000002l	undefined	23304
	utg000006l	undefined	618933
	utg000007l	undefined	122221
	utg000009l	undefined	50190
	utg000010l	undefined	18666
NCTC10988	utg000001l	chromosomal	470745
	utg000002l	chromosomal	1143622
	utg000003l	chromosomal	39170
	utg000004l	chromosomal	1521633
	utg000005l	chromosomal	35669
	utg000006l	chromosomal	36182
	utg000007l	chromosomal	28025
	utg000009l	chromosomal	23011
	utg000011l	chromosomal	25778
	total chromosomal length		3323835
	utg000010c	undefined	27255
	utg000008c	none	1813
NCTC11126	utg000001l	chromosomal	799550
	utg000003l	chromosomal	39468
	utg000005l	chromosomal	199017
	utg000006l	chromosomal	654158
	utg000007l	chromosomal	801856
	utg000008l	chromosomal	150048
	utg000009l	chromosomal	615460
	utg000010l	chromosomal	446084
	utg000012l	chromosomal	187190
	utg000013l	chromosomal	24165
	utg000015l	chromosomal	124049
	utg000016l	chromosomal	115589
	utg000017l	chromosomal	214509
	utg000018l	chromosomal	36740
	utg000019l	chromosomal	18016
	total chromosomal length		4425899
	utg000002l	undefined	463729
	utg000004l	none	129544
	utg000011l	none	152040
	utg000014l	none	20861
NCTC11343	utg000001l	chromosomal	1137077
	utg000008l	chromosomal	387994
	utg000009l	chromosomal	303645
	utg000013l	chromosomal	73842
	utg000017l	chromosomal	44820
	utg000021l	chromosomal	71812
	utg000023l	chromosomal	82330
	utg000026l	chromosomal	37829
	utg000027l	chromosomal	20725
	utg000028l	chromosomal	10169
	total chromosomal length		2170243
	utg000002l	undefined	110639
	utg000003l	undefined	629323
	utg000004l	undefined	483485
	utg000005l	undefined	88713
	utg000006l	undefined	94951
	utg000007l	undefined	268271
	utg000010l	undefined	265843
	utg000011l	undefined	186796
	utg000012l	undefined	244669
	utg000014l	undefined	739847
	utg000015l	undefined	328310
	utg000016l	undefined	159212
	utg000018l	undefined	95506
	utg000019l	undefined	84352
	utg000020l	undefined	49752
	utg000022l	undefined	78366
	utg000024l	undefined	25887
	utg000025l	undefined	63812
NCTC11360	utg000001l	chromosomal	83040
	utg000002l	chromosomal	142687
	utg000003l	chromosomal	86224
	utg000004l	chromosomal	117971
	utg000005l	chromosomal	103040
	utg000006l	chromosomal	379750
	utg000007l	chromosomal	73713
	utg000008l	chromosomal	87575
	utg000009l	chromosomal	173405
	utg000010l	chromosomal	39660

	utg0000111	chromosomal	75956
	utg0000121	chromosomal	43377
	utg0000131	chromosomal	104822
	utg0000141	chromosomal	66226
	utg0000151	chromosomal	97577
	utg0000161	chromosomal	22672
	utg0000171	chromosomal	68226
	utg0000181	chromosomal	31089
	utg0000191	chromosomal	135026
	utg0000201	chromosomal	12813
	utg0000211	chromosomal	59788
	utg0000221	chromosomal	18821
	utg0000231	chromosomal	10147
	utg0000241	chromosomal	14492
	utg0000251	chromosomal	12493
	total chromosomal length		2060590
NCTC11435	utg0000011	chromosomal	348162
	utg000002c	chromosomal	1514816
	utg0000031	chromosomal	992029
	utg0000041	chromosomal	641762
	utg0000051	chromosomal	861119
	utg0000061	chromosomal	287740
	total chromosomal length		4645628
	utg000007c	none	1992
NCTC11800	utg0000011	chromosomal	3251923
	utg0000021	chromosomal	430158
	utg0000031	chromosomal	887997
	total chromosomal length		4570078
	utg0000041	undefined	190071
NCTC11872	utg0000011	chromosomal	171196
	utg0000021	chromosomal	82073
	utg0000031	chromosomal	411282
	utg0000041	chromosomal	188111
	utg0000051	chromosomal	116854
	utg0000061	chromosomal	68409
	utg0000071	chromosomal	132209
	utg0000081	chromosomal	135142
	utg0000091	chromosomal	198320
	utg0000101	chromosomal	170105
	utg0000111	chromosomal	62719
	utg0000121	chromosomal	193447
	utg0000131	chromosomal	33052
	total chromosomal length		1962919
NCTC12123	utg0000011	chromosomal	2853675
	utg0000021	chromosomal	2105813
	utg0000031	chromosomal	40089
	total chromosomal length		4999577
	utg000005c	plasmidic	10039
	utg0000041	undefined	31948
NCTC12126	utg0000011	chromosomal	319898
	utg0000021	chromosomal	731428
	utg0000031	chromosomal	2015098
	utg0000041	chromosomal	424548
	utg0000051	chromosomal	234649
	utg0000061	chromosomal	95263
	utg0000071	chromosomal	401716
	utg0000081	chromosomal	165317
	utg0000091	chromosomal	79334
	utg0000101	chromosomal	121403
	utg0000111	chromosomal	117059
	utg0000131	chromosomal	100109
	utg0000141	chromosomal	144200
	utg0000151	chromosomal	64609
	utg0000161	chromosomal	59101
	total chromosomal length		5073732
	utg0000121	undefined	169903
NCTC12131	utg0000011	chromosomal	111790
	utg0000021	chromosomal	92827
	utg0000031	chromosomal	97030
	utg0000041	chromosomal	124270
	utg0000051	chromosomal	69426
	utg0000071	chromosomal	26064
	utg0000081	chromosomal	39017
	utg0000091	chromosomal	161140
	utg0000101	chromosomal	67339
	utg0000131	chromosomal	30391
	utg0000141	chromosomal	100618
	utg0000151	chromosomal	125621
	utg0000161	chromosomal	60566
	utg0000171	chromosomal	16554
	utg0000181	chromosomal	195541
	utg0000191	chromosomal	101170
	utg0000201	chromosomal	130409
	utg0000211	chromosomal	49770
	utg0000221	chromosomal	84913
	utg0000231	chromosomal	78770
	utg0000241	chromosomal	257054
	utg0000251	chromosomal	154990
	utg0000261	chromosomal	38811
	utg0000271	chromosomal	68728
	utg0000281	chromosomal	46799
	utg0000301	chromosomal	161317
	utg0000311	chromosomal	71407
	utg0000321	chromosomal	77866
	utg0000331	chromosomal	86540
	utg0000341	chromosomal	114741
	utg0000361	chromosomal	60369
	utg0000371	chromosomal	28783
	utg0000381	chromosomal	71999
	utg0000391	chromosomal	64473
	utg0000401	chromosomal	72097
	utg0000411	chromosomal	12003
	utg0000421	chromosomal	19063
	utg0000431	chromosomal	29043
	utg0000441	chromosomal	36339
	utg0000451	chromosomal	134261
	utg0000461	chromosomal	56314
	utg0000471	chromosomal	101727
	utg0000481	chromosomal	16300
	utg0000491	chromosomal	101615
	utg0000501	chromosomal	25187
	utg0000511	chromosomal	23598
	utg0000521	chromosomal	17725
	utg0000531	chromosomal	106162
	utg0000541	chromosomal	18067
	utg0000551	chromosomal	87197
	utg0000561	chromosomal	38172
	utg0000571	chromosomal	68391
	utg0000581	chromosomal	38284
	utg0000611	chromosomal	73927
	utg0000621	chromosomal	34241
	utg0000631	chromosomal	30728
	utg0000641	chromosomal	22173
	utg0000651	chromosomal	15783
	utg0000661	chromosomal	25123
	utg0000671	chromosomal	49646
	utg0000681	chromosomal	15064
	utg0000691	chromosomal	38231
	utg0000701	chromosomal	25850
	utg0000721	chromosomal	103746
	utg0000731	chromosomal	32616
	utg0000741	chromosomal	20672
	utg0000751	chromosomal	20049
	utg0000761	chromosomal	16489
	utg0000781	chromosomal	7701
	utg0000791	chromosomal	17282
	utg0000801	chromosomal	25386
	utg0000811	chromosomal	13893
	utg0000821	chromosomal	21851
	utg0000841	chromosomal	6442
	utg0000851	chromosomal	4765
	utg0000861	chromosomal	8500

	utg0000871	chromosomal	5363
	total chromosomal length		4704169
	utg0000061	undefined	14352
	utg0000111	undefined	40323
	utg0000121	undefined	26155
	utg0000291	undefined	24896
	utg0000591	undefined	21237
	utg0000601	undefined	5107
	utg0000711	undefined	44778
	utg0000771	undefined	11705
	utg0000831	undefined	9380
	utg0000351	none	65396
NCTC12132	utg0000011	chromosomal	2895268
	utg000002c	chromosomal	536810
	utg0000031	chromosomal	69541
	utg0000041	chromosomal	19313
	total chromosomal length		3520932
NCTC12146	utg0000011	chromosomal	5930232
	total chromosomal length		5930232
NCTC12694	utg0000011	chromosomal	58159
	utg0000021	chromosomal	35524
	utg0000031	chromosomal	28409
	utg0000041	chromosomal	16449
	utg0000051	chromosomal	17394
	utg0000061	chromosomal	21530
	utg0000071	chromosomal	11853
	utg0000081	chromosomal	8442
	utg0000091	chromosomal	10039
	utg0000101	chromosomal	49546
	utg0000111	chromosomal	26813
	utg0000121	chromosomal	23265
	utg0000131	chromosomal	21153
	utg0000141	chromosomal	12511
	utg0000151	chromosomal	17131
	utg0000161	chromosomal	42769
	utg0000171	chromosomal	12766
	utg0000181	chromosomal	24545
	utg0000191	chromosomal	23162
	utg0000201	chromosomal	19756
	utg0000211	chromosomal	45548
	utg0000221	chromosomal	21115
	utg0000231	chromosomal	25591
	utg0000241	chromosomal	30239
	utg0000251	chromosomal	21095
	utg0000261	chromosomal	9863
	utg0000271	chromosomal	23159
	utg0000291	chromosomal	8102
	utg0000301	chromosomal	11082
	utg0000311	chromosomal	28054
	utg0000331	chromosomal	42654
	utg0000341	chromosomal	30933
	utg0000351	chromosomal	46346
	utg0000361	chromosomal	30903
	utg0000371	chromosomal	25189
	utg0000381	chromosomal	18193
	utg0000391	chromosomal	43343
	utg0000401	chromosomal	30397
	utg0000411	chromosomal	15935
	utg0000421	chromosomal	19588
	utg0000431	chromosomal	18820
	utg0000441	chromosomal	23764
	utg0000451	chromosomal	27579
	utg0000461	chromosomal	26394
	utg0000471	chromosomal	18649
	utg0000481	chromosomal	20699
	utg0000491	chromosomal	28188
	utg0000501	chromosomal	7493
	utg0000511	chromosomal	46315
	utg0000531	chromosomal	5107
	utg0000541	chromosomal	40004
	utg0000551	chromosomal	20811
	utg0000561	chromosomal	6201
	utg0000571	chromosomal	42503
	utg0000581	chromosomal	9998
	utg0000591	chromosomal	14321
	utg0000601	chromosomal	22892
	utg0000611	chromosomal	28867
	utg0000621	chromosomal	28898
	utg0000641	chromosomal	24612
	utg0000651	chromosomal	41956
	utg0000661	chromosomal	22542
	utg0000671	chromosomal	18013
	utg0000681	chromosomal	30693
	utg0000691	chromosomal	20647
	utg0000701	chromosomal	35946
	utg0000711	chromosomal	26657
	utg0000721	chromosomal	18854
	utg0000731	chromosomal	8301
	utg0000741	chromosomal	6505
	utg0000751	chromosomal	18941
	utg0000761	chromosomal	17232
	utg0000771	chromosomal	21805
	utg0000781	chromosomal	14634
	utg0000791	chromosomal	37447
	utg0000801	chromosomal	12665
	utg0000811	chromosomal	39494
	utg0000821	chromosomal	17950
	utg0000831	chromosomal	21934
	utg0000841	chromosomal	22518
	utg0000851	chromosomal	7041
	utg0000861	chromosomal	28654
	utg0000871	chromosomal	16884
	utg0000881	chromosomal	28723
	utg0000891	chromosomal	16734
	utg0000901	chromosomal	13996
	utg0000911	chromosomal	23988
	utg0000921	chromosomal	14611
	utg0000931	chromosomal	7501
	utg0000941	chromosomal	25459
	utg0000951	chromosomal	4562
	utg0000961	chromosomal	47136
	utg0000971	chromosomal	4814
	utg0000981	chromosomal	25935
	utg0000991	chromosomal	10951
	utg0001001	chromosomal	21525
	utg0001011	chromosomal	10732
	utg0001021	chromosomal	6168
	utg0001031	chromosomal	15062
	utg0001041	chromosomal	11927
	utg0001051	chromosomal	10715
	utg0001061	chromosomal	7331
	utg0001071	chromosomal	10881
	utg0001081	chromosomal	15574
	utg0001091	chromosomal	27587
	utg0001101	chromosomal	23469
	utg0001111	chromosomal	12398
	utg0001121	chromosomal	57852
	utg0001131	chromosomal	13429
	utg0001141	chromosomal	15167
	utg0001151	chromosomal	29785
	utg0001161	chromosomal	21769
	utg0001171	chromosomal	8477
	utg0001181	chromosomal	16251
	utg0001191	chromosomal	21816
	utg0001201	chromosomal	4937
	utg0001211	chromosomal	15924
	utg0001221	chromosomal	14350
	utg0001231	chromosomal	17284
	utg0001241	chromosomal	14773
	utg0001251	chromosomal	13950
	utg0001261	chromosomal	12821
	utg0001271	chromosomal	12776

	total chromosomal length		2667113
	utg000028l	plasmidic	39729
	utg000063l	plasmidic	24889
	utg000032l	none	10367
	utg000052l	none	44932
NCTC12841	utg000001l	chromosomal	2025517
	total chromosomal length		2025517
NCTC12993	utg000001l	chromosomal	4585710
	utg000002l	chromosomal	728592
	total chromosomal length		5314302
	utg000004l	plasmidic	83189
	utg000006l	plasmidic	25452
	utg000007l	plasmidic	6215
	utg000003l	undefined	109233
	utg000005l	undefined	74705
NCTC12998	utg000001l	chromosomal	669883
	utg000002l	chromosomal	3936078
	utg000003l	chromosomal	584057
	utg000004l	chromosomal	737303
	total chromosomal length		5927321
	utg000005c	plasmidic	125518
NCTC13095	utg000001c	chromosomal	5922307
	utg000003l	chromosomal	34891
	total chromosomal length		5957198
	utg000002c	undefined	117186
	utg000004c	undefined	164749
NCTC13125	utg000001l	chromosomal	4270665
	utg000002l	chromosomal	470580
	utg000003l	chromosomal	287719
	utg000004l	chromosomal	62610
	utg000005l	chromosomal	392905
	utg000006l	chromosomal	215108
	utg000009l	chromosomal	25575
	total chromosomal length		5725162
	utg000007c	plasmidic	105857
	utg000008c	plasmidic	93624
	utg000010c	undefined	41914
NCTC13348	utg000001l	chromosomal	297750
	utg000002l	chromosomal	152446
	utg000003l	chromosomal	219639
	utg000004l	chromosomal	237470
	utg000006l	chromosomal	471362
	utg000007l	chromosomal	399972
	utg000008l	chromosomal	226216
	utg000009l	chromosomal	979865
	utg000010l	chromosomal	154813
	utg000011l	chromosomal	97347
	utg000012l	chromosomal	283827
	utg000013l	chromosomal	408789
	utg000014l	chromosomal	469421
	utg000015l	chromosomal	431950
	utg000016l	chromosomal	157498
	utg000017l	chromosomal	129726
	utg000018l	chromosomal	18621
	total chromosomal length		5136712
	utg000005l	plasmidic	105592
NCTC13463	utg000001l	chromosomal	2602844
	utg000002l	chromosomal	476415
	utg000003l	chromosomal	2106813
	utg000006l	chromosomal	151619
	total chromosomal length		5337691
	utg000004c	plasmidic	89940
	utg000005c	undefined	50683
	utg000007l	undefined	2881
NCTC13543	utg000001c	chromosomal	3006943
	utg000002l	chromosomal	2166302
	utg000007l	chromosomal	54242
	total chromosomal length		5227487
	utg000005l	plasmidic	24866
	utg000006l	plasmidic	57848
	utg000008l	plasmidic	14511
	utg000004c	undefined	509202
	utg000009l	undefined	15019
	utg000003l	none	163283
NCTC4672	utg000001l	chromosomal	390339
	utg000002l	chromosomal	234263
	utg000003l	chromosomal	205142
	utg000004l	chromosomal	194700
	utg000005l	chromosomal	143175
	utg000006l	chromosomal	145320
	utg000007l	chromosomal	20463
	utg000008l	chromosomal	72927
	utg000009l	chromosomal	171587
	utg000010l	chromosomal	152272
	utg000011l	chromosomal	39361
	utg000012l	chromosomal	68994
	utg000013l	chromosomal	26376
	utg000014l	chromosomal	41255
	utg000015l	chromosomal	60948
	utg000016l	chromosomal	13117
	total chromosomal length		1980239
NCTC5050	utg000001l	chromosomal	1602894
	utg000002l	chromosomal	1353385
	utg000003l	chromosomal	1365254
	utg000004l	chromosomal	1153772
	total chromosomal length		5475305
	utg000005c	plasmidic	116760
	utg000006c	plasmidic	40870
	utg000007c	plasmidic	76667
	utg000008l	none	37817
NCTC5053	utg000001l	chromosomal	1215813
	utg000002l	chromosomal	334789
	utg000004l	chromosomal	1438231
	utg000006l	chromosomal	675061
	utg000007l	chromosomal	225750
	utg000008l	chromosomal	372066
	utg000009l	chromosomal	188575
	utg000010l	chromosomal	677706
	utg000011l	chromosomal	244959
	utg000012l	chromosomal	63154
	utg000013l	chromosomal	26449
	total chromosomal length		5462553
	utg000005c	plasmidic	208263
	utg000014c	plasmidic	102763
	utg000015c	plasmidic	104579
	utg000016c	plasmidic	12712
	utg000018c	plasmidic	5167
	utg000003c	undefined	105118
	utg000017c	none	2115
NCTC5055	utg000001l	chromosomal	5214489
	utg000002l	chromosomal	13329
	utg000004l	chromosomal	16397
	utg000006l	chromosomal	11993
	utg000007l	chromosomal	11530
	utg000008l	chromosomal	23643
	utg000010l	chromosomal	12125
	utg000011l	chromosomal	20225
	utg000012l	chromosomal	11167
	utg000013l	chromosomal	10465
	utg000014l	chromosomal	11390
	utg000015l	chromosomal	21429
	utg000016l	chromosomal	11492
	utg000017l	chromosomal	14583
	utg000018l	chromosomal	10242
	utg000019l	chromosomal	20037
	utg000020l	chromosomal	12298
	utg000022l	chromosomal	21719
	utg000023l	chromosomal	20549
	utg000024l	chromosomal	7808
	total chromosomal length		5496910
	utg000003l	plasmidic	19472
	utg000009l	plasmidic	34026

	utg000021l	plasmidic	17587
	utg000005l	undefined	10762
NCTC7152	utg000001l	chromosomal	4797250
	utg000003l	chromosomal	446534
	total chromosomal length		5243784
	utg000002c	plasmidic	140333
	utg000004l	plasmidic	45336
NCTC7922	utg000001l	chromosomal	223615
	utg000002l	chromosomal	3332503
	utg000003l	chromosomal	633327
	utg000004l	chromosomal	213700
	utg000005l	chromosomal	47119
	utg000006l	chromosomal	541108
	utg000007l	chromosomal	353127
	utg000010l	chromosomal	56267
	utg000011l	chromosomal	15478
	total chromosomal length		5416244
	utg000008c	plasmidic	84907
	utg000009c	plasmidic	59289
	utg000012c	undefined	5073
	utg000013c	undefined	5631
	utg000014c	none	1251
NCTC8179	utg000001l	chromosomal	865090
	utg000002l	chromosomal	1960752
	utg000003l	chromosomal	277217
	utg000004l	chromosomal	786279
	utg000005l	chromosomal	111098
	utg000006l	chromosomal	93197
	utg000007l	chromosomal	293360
	utg000008l	chromosomal	24186
	utg000010l	chromosomal	292276
	utg000012l	chromosomal	125358
	utg000013l	chromosomal	265142
	utg000014l	chromosomal	277553
	utg000015l	chromosomal	294414
	utg000016l	chromosomal	32286
	utg000017l	chromosomal	22000
	total chromosomal length		5720208
	utg000009c	plasmidic	118435
	utg000011c	plasmidic	71383
	utg000018c	none	3105
NCTC8500	utg000001c	chromosomal	4831304
	total chromosomal length		4831304
	utg000002c	plasmidic	55643
NCTC8684	utg000001l	chromosomal	1077377
	utg000006l	chromosomal	1059624
	total chromosomal length		2137001
	utg000002l	undefined	750049
	utg000003l	undefined	701800
	utg000004l	undefined	762169
	utg000005l	undefined	622781
	utg000007l	undefined	49586
NCTC9075	utg000001l	chromosomal	2460452
	utg000002l	chromosomal	3016645
	utg000004l	chromosomal	9967
	total chromosomal length		5487064
	utg000003c	undefined	52022
NCTC9078	utg000001c	chromosomal	4242489
	utg000002c	chromosomal	1146428
	total chromosomal length		5388917
	utg000004c	plasmidic	87254
	utg000003c	none	132921
NCTC9098	utg000001l	chromosomal	3054707
	utg000002l	chromosomal	1563634
	utg000003l	chromosomal	219418
	utg000005l	chromosomal	337874
	utg000007l	chromosomal	29947
	utg000008l	chromosomal	288339
	total chromosomal length		5493919
	utg000004c	plasmidic	52547
	utg000006c	plasmidic	88507
	utg000009c	none	3685
	utg000010c	none	2517
NCTC9111	utg000002l	chromosomal	382296
	utg000003l	chromosomal	258361
	utg000004l	chromosomal	1397867
	utg000005l	chromosomal	120406
	utg000006l	chromosomal	360884
	utg000007l	chromosomal	354979
	utg000008l	chromosomal	393118
	utg000009l	chromosomal	780946
	utg000010l	chromosomal	496328
	utg000011c	chromosomal	751321
	utg000013l	chromosomal	14110
	utg000014l	chromosomal	298974
	utg000018c	chromosomal	9151
	total chromosomal length		5618741
	utg000001c	plasmidic	126297
	utg000012c	plasmidic	92083
	utg000015c	plasmidic	71898
	utg000016c	undefined	5475
	utg000017c	undefined	97293
NCTC9112	utg000001l	chromosomal	267187
	utg000002l	chromosomal	38684
	utg000003l	chromosomal	799223
	utg000004l	chromosomal	226433
	utg000005l	chromosomal	1445001
	utg000006l	chromosomal	583111
	utg000007l	chromosomal	772601
	utg000008l	chromosomal	739996
	utg000009l	chromosomal	283544
	utg000010l	chromosomal	537734
	utg000011l	chromosomal	134545
	utg000012l	chromosomal	20182
	utg000013l	chromosomal	22078
	utg000014l	chromosomal	19079
	utg000015l	chromosomal	17055
	total chromosomal length		5906453
	utg000016c	none	786
NCTC9184	utg000001l	chromosomal	29829
	utg000002l	chromosomal	18494
	utg000004l	chromosomal	16159
	utg000005l	chromosomal	24250
	utg000006l	chromosomal	18494
	utg000007l	chromosomal	21484
	utg000008l	chromosomal	27067
	utg000009l	chromosomal	16993
	utg000010l	chromosomal	23540
	utg000011l	chromosomal	22732
	utg000012l	chromosomal	6381
	utg000013l	chromosomal	29998
	utg000014l	chromosomal	22501
	utg000015l	chromosomal	22166
	utg000016l	chromosomal	14255
	utg000017l	chromosomal	13452
	utg000018l	chromosomal	19121
	total chromosomal length		346916
	utg000003c	none	4242
NCTC9645	utg000001l	chromosomal	43611
	utg000002l	chromosomal	159710
	utg000003l	chromosomal	60098
	utg000004l	chromosomal	148353
	utg000005l	chromosomal	202718
	utg000006l	chromosomal	19551
	utg000007l	chromosomal	50893
	utg000008l	chromosomal	131157
	utg000009l	chromosomal	154075
	utg000010l	chromosomal	213751
	utg000011l	chromosomal	29643
	utg000012l	chromosomal	39758
	utg000013l	chromosomal	79432
	utg000014l	chromosomal	47565

utg000016l	chromosomal	24983
utg000017l	chromosomal	60388
utg000018l	chromosomal	36168
utg000019l	chromosomal	96948
utg000020l	chromosomal	163752
utg000021l	chromosomal	90608
utg000022l	chromosomal	56793
utg000023l	chromosomal	182187
utg000024l	chromosomal	51393
utg000025l	chromosomal	53299
utg000026l	chromosomal	128083
utg000027l	chromosomal	124407
utg000028l	chromosomal	48273
utg000029l	chromosomal	126118
utg000030l	chromosomal	43482
utg000032l	chromosomal	169005
utg000033l	chromosomal	48040
utg000034l	chromosomal	28371
utg000035l	chromosomal	54647
utg000036l	chromosomal	17500
utg000037l	chromosomal	99129
utg000038l	chromosomal	109018
utg000039l	chromosomal	169360
utg000040l	chromosomal	79322
utg000041l	chromosomal	29162
utg000042l	chromosomal	12864
utg000043l	chromosomal	135445
utg000044l	chromosomal	77604
utg000045l	chromosomal	212393
utg000046l	chromosomal	53152
utg000047l	chromosomal	23857
utg000048l	chromosomal	87995
utg000050l	chromosomal	8353
utg000053l	chromosomal	61272
utg000054l	chromosomal	19953
utg000055l	chromosomal	76527
utg000056l	chromosomal	30382
utg000057l	chromosomal	38221
utg000058l	chromosomal	67870
utg000059l	chromosomal	33128
utg000060l	chromosomal	12588
utg000061l	chromosomal	93957
utg000062l	chromosomal	37019
utg000063l	chromosomal	24109
utg000064l	chromosomal	68089
utg000066l	chromosomal	39231
utg000067l	chromosomal	53886
utg000068l	chromosomal	49692
utg000069l	chromosomal	69014
utg000070l	chromosomal	15069
utg000071l	chromosomal	12499
utg000072l	chromosomal	34026
utg000073l	chromosomal	33264
utg000074l	chromosomal	11225
utg000075l	chromosomal	75997
utg000077l	chromosomal	23613
utg000078l	chromosomal	18064
utg000079l	chromosomal	13912
utg000080l	chromosomal	20491
utg000081l	chromosomal	19414
utg000082l	chromosomal	17026
utg000084l	chromosomal	10403
total chromosomal length		5162355
utg000015l	plasmidic	289749
utg000031c	plasmidic	85333
utg000052l	plasmidic	89214
utg000083l	plasmidic	9793
utg000065l	undefined	27749
utg000076l	undefined	14407
utg000049l	none	20985
utg000051l	none	21105

NCTC9646	utg000001l	chromosomal	1824786
	utg000002l	chromosomal	1185729
	utg000004l	chromosomal	918891
	utg000005l	chromosomal	1259981
	utg000006l	chromosomal	169448
	utg000007l	chromosomal	22892
	utg000008l	chromosomal	205862
	utg000009l	chromosomal	61801
	utg000010l	chromosomal	105678
	total chromosomal length		5755068
	utg000003l	plasmidic	181111
	utg000011l	plasmidic	38663
	utg000012l	plasmidic	16804
	utg000015l	plasmidic	26818
	utg000016l	plasmidic	6887
	utg000014l	undefined	37740
	utg000013c	none	2141
NCTC9695	utg000001l	chromosomal	4677804
	total chromosomal length		4677804
	utg000002c	none	3416

Table 7: `Miniasm` contigs classification per dataset. Total length of chromosomal contigs is given.

Appendix 7 Assembly length

In Table 8 we report the sum of lengths of all contigs in assemblies computed by `Miniasm`, `Canu`, and the assemblies produced by NCTC.

Dataset	`Canu`	`Miniasm`	NCTC
NCTC10006	5297110	5586964	5285365
NCTC10332	6353743	6510723	6316979
NCTC10444	5358052	5575543	5295042
NCTC10702	3140906	3113949	3044394
NCTC10766	5704549	5981360	5662808
NCTC10794	2323585	2220964	2164041
NCTC10988	3242798	3352903	3309451
NCTC11126	4887988	5192073	4875981
NCTC11343	6102368	6167977	5984896
NCTC11360	2189222	2060590	2078787
NCTC11435	4452377	4647620	4443087
NCTC11800	4482172	4760149	4461490
NCTC11872	1881379	1962919	1879445
NCTC12123	4889243	5041564	4785686
NCTC12126	5037863	5243635	5015169
NCTC12131	4799906	4967498	4743059
NCTC12132	3360769	3520932	3326136
NCTC12146	5648936	5930232	5621322
NCTC12694	4667053	2787030	4439218
NCTC12841	1998859	2025517	*
NCTC12993	5339609	5613096	5287156
NCTC12998	5754007	6052839	5723058
NCTC13095	6018682	6239133	5962730
NCTC13125	5868476	5966557	3814513
NCTC13348	5133106	5242304	5042742
NCTC13463	5298325	5481195	5268825
NCTC13543	5834577	6012216	5822755
NCTC4672	2113924	1980239	1942171
NCTC5050	5360294	5747419	5342107
NCTC5053	5838871	6003270	5769583
NCTC5055	5618297	5578757	4924715
NCTC7152	5136487	5429453	5086417
NCTC7922	5497660	5572395	5367566
NCTC8179	5979700	5913131	5715723
NCTC8500	4733397	4886947	4709652
NCTC8684	4936541	5023386	4860337
NCTC9075	5336449	5539086	5322538
NCTC9078	5458474	5609092	5430557
NCTC9098	5385311	5641175	5331923
NCTC9111	5942101	6011787	5859950
NCTC9112	5557088	5907239	5468741
NCTC9184	2541470	351158	679813
NCTC9645	5930294	5720690	5852841
NCTC9646	5874126	6065232	*
NCTC9695	4792855	4681220	4738566

Table 8: Total length of chromosomal contigs for `Canu` and `Miniasm` assemblers, total length of all contigs for NCTC on the 45 NCTC datasets studied. "*" means that NCTC assembly is not available.

Appendix 8 Detailed assembly results

Supplementary *T. roseus* figures

Total number of reads	11592
Total length	104,608,900bp
Longest reads	46,221bp
Shortest reads	4bp
Mean Length	9,024bp
Median Length	6,978bp
N10	364 reads
N50	2586 reads
N90	7236 reads
L10	24,917bp
L50	14,590bp
L90	4,550bp

Table 9: Some statistics about reads produced by LongISLND for *T. roseus* synthetic dataset.

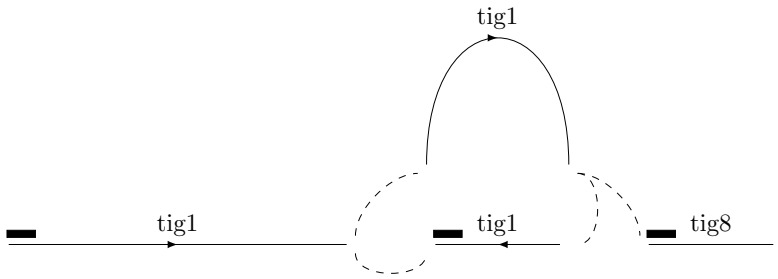

Figure 2: Canu *T. roseus* tig1 reconstruction. Plain lines denote path without branches in the SG (the one shown Figure 2b). Boxes denote reads at contig extremities. Dashed lines denote overlaps between reads (and then contig extremities). Arrows show the path used by Canu to build tig1: it goes through the "loop" before going back.

Appendix 9 Supplementary NCTC figures

NCTC12123

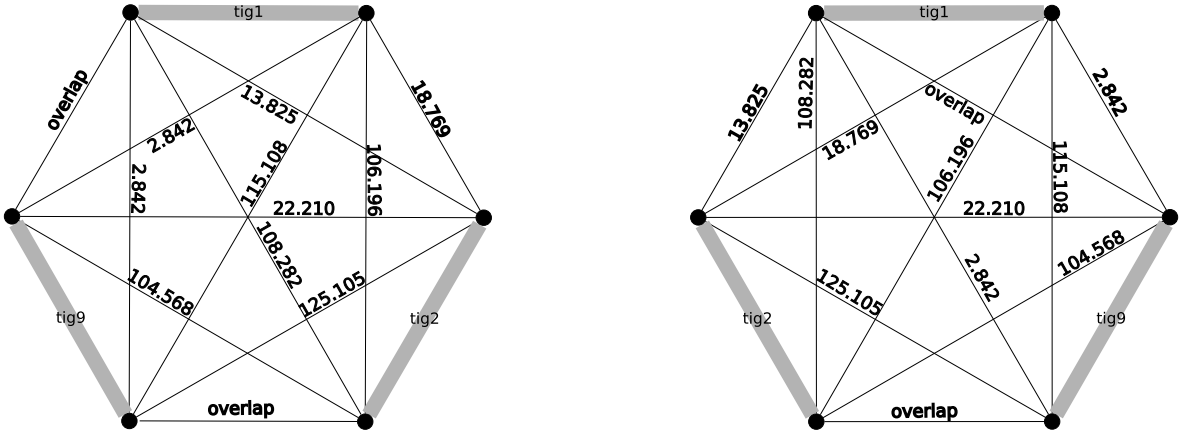

Figure 3: Shortest paths in AAG. Both scenarii (paths that follow edges with bold label) use the edge of weight 7.178, the only remaining ambiguity concerns the order of tig1 against the pair tig2/tig9. 'overlap' indicate than our pipeline found an overlap between the contig extremities. The left scenario has a weight of 29.379 ($22.201 + 7.178$) while the right one has a weight of 30.736 ($17.209 + 7.178 + 6.349$).

NCTC5050

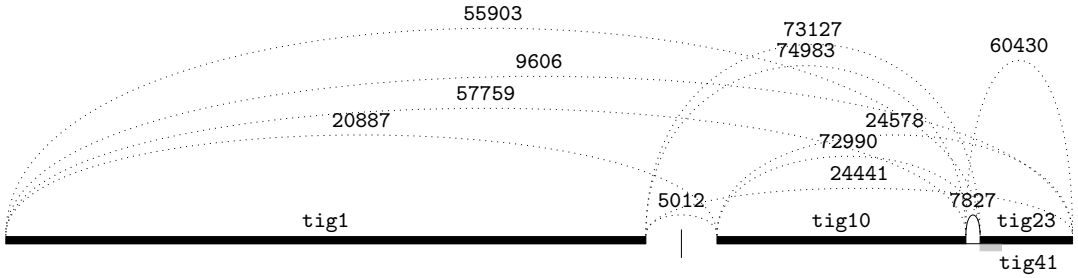

Figure 4: NCTC5050 contigs mapped against NCTC reference for ordering. Paths are shown along with their number of bases as labels. The NCTC assembly consists of two contigs, hence the relative order of tig1 and tig10/tig23 cannot be inferred (vertical line in the figure). We observed that a portion of tig1 is inverted with respect to the NCTC assembly, with no impact on the path analysis as this putative misassembly does not involve an extremity of the contig.

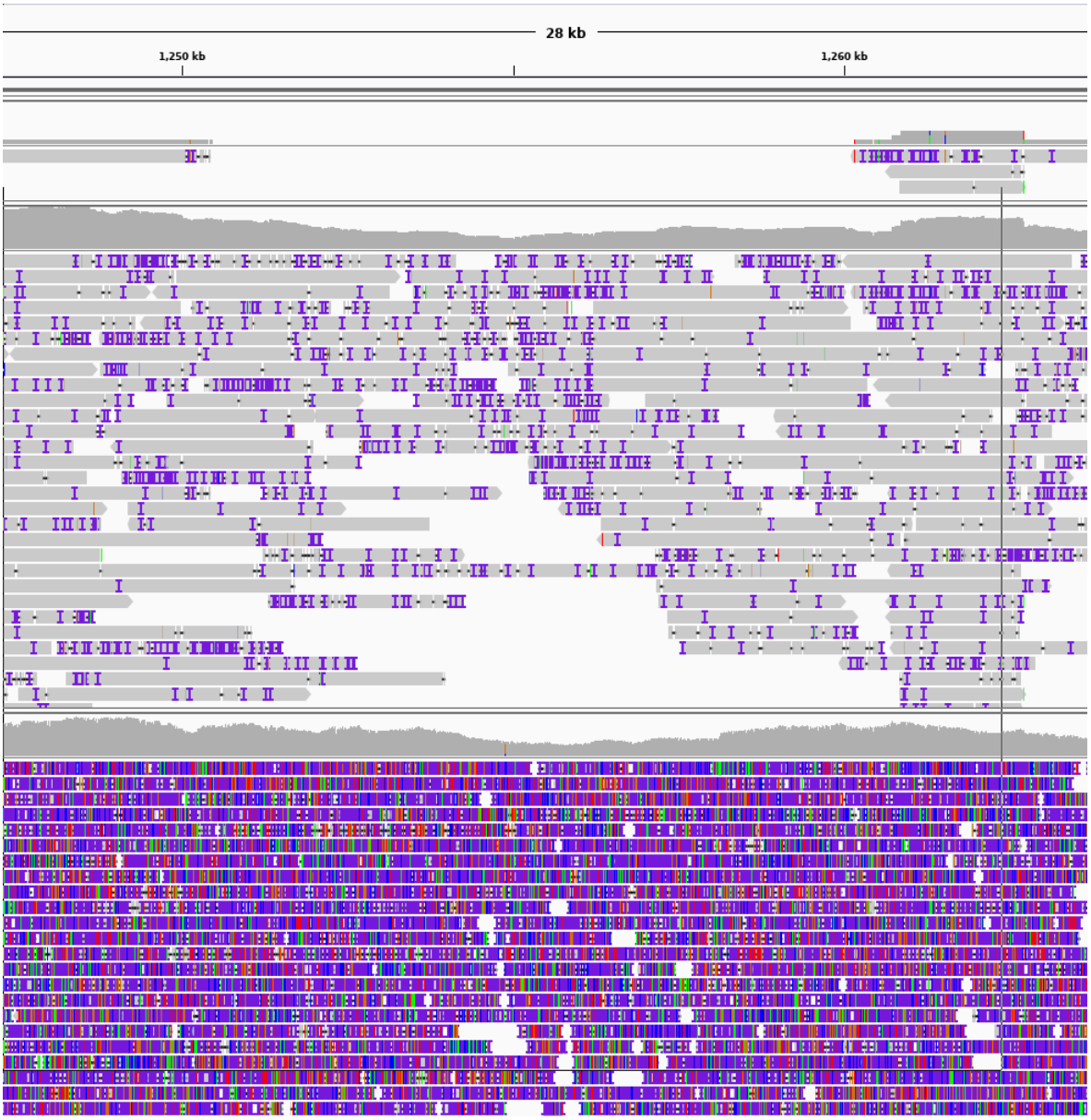

Figure 5: IGV view of NCTC5050 mapping of `Canu` contig against NCTC contig, in junction between tig10 (first track at left) and tig23 (first track at right), tig41 are mapped on begin of tig23 in forward and reverse. The second track represent the mapping of `Canu` corrected read, the third track represent the raw reads. Above each this track we can observe the coverage curve and drop of this curve between the tig10 and tig23, for corrected read is around 50x coverage before junction, equal to 15x at minimal, and less than 40x after junction, this value are 90x, 25x and 40x for raw read. In addition we can observe more error in corrected read on this drop of coverage.

Appendix 10 YACRD: Yet Another Chimeric Read Detector

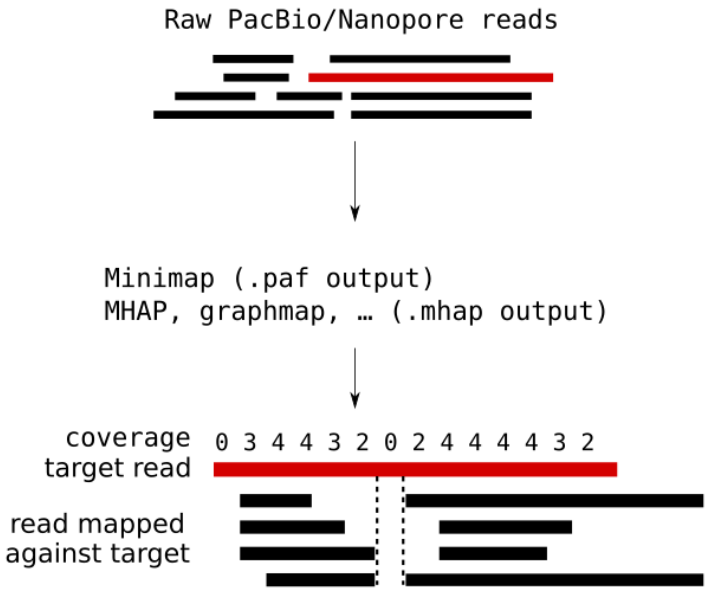

Figure 6: YACRD (manuscript in preparation) detects chimeric regions present in the read dataset. To detect such regions, YACRD takes as input the output of an overlapper (both PAF and MHAP format are accepted). For each read in the dataset, YACRD computes positional coverage values based on the overlaps with that read. If there is a drop of coverage, the corresponding read is marked as 'chimeric'.