Genomics applications of nanopore long-read sequencing for small to large sized genomes

Liem, M.

— Chapter 3

Genome assembly of the transposon-enriched *Allorhizobium* strain LBA9072

Chapter in preparation for publication

— Abstract

The assembly of a transposon-enriched in-house *Allorhizobium* strain was improved from 154 contigs to two circular chromosomes and two additional plasmids using Oxford Nanopore Technologies (ONT) long-reads. We have assembled the sequencing data using assemblers Unicycler, Flye and Canu, using hybrid and *de novo* assembly strategies. Assembly differences are specifically apparent for the Canu assembly, where the large chromosome was still separated into two distinct contigs and unable to circularize a plasmid. Both hybrid and *de novo* assembly results show high sequence similarity compared to a reference, although some misassemblies are found within the Canu assembly. The frequency and location of two prominently present transposons are identified in addition to transposable elements found by the ISfinder tool. The lack of sequence similarity between the reference and the final assembly around transposon locations suggest that genomic diversification is facilitated by transposons.

— Introduction

Here we investigate the genomic structure of an in-house *Allorhizobium* strain (LBA9072), recently reclassified and previously considered an *Agrobacterium* strain[1]. Bacteria of the genus *Agrobacterium* are soil-borne plant pathogens that cause crown gall disease and are also used extensively in genetic engineering[2,3]. Its genome is usually of moderate complexity, consisting of 2 chromosomes (both circular), a tumor-inducing plasmid (pTi, usually ~200 Kb) and a larger 'cryptic plasmid' of unknown function, and sometimes additional plasmids[4-6]. This pathogen can transfer and integrate a part of its genome (the tumor-inducing T-DNA, found on pTi) into a plant host, reprogramming cells to a proliferation state/ phenotype and resulting in plant tumor formation. Gene expression patterns of host plants show different characteristics which depend on bacterium strain, specialization of strains, plant species and infected cell-type[7]. The genomic structure of *Allorhizobium* is comparable however it contains two circular chromosomes. The genomic structure of the strain investigated in this study deviates from most well-studied strains since it contains large numbers of transposable elements that occur at multiple locations throughout the genome. The lengths of those transposons are longer that Illumina sequencing reads, hence preventing whole genome assembly using Illumina data alone. Since the genomic structure of bacterial genomes is highly versatile it is important to investigating the genomic structure of individual strains. In this study we investigate in detail how long-read sequencing data enables the assembly of complex genomic content, potentially resulting in chromosome-scale contigs.

— Materials and method

Illumina sequencing and Velvet assembly on Illumina data

We have generated 99-nucleotide paired-end reads on Illumina HiSeq with 150x coverage for our strain LBA9072 and used Velvet (version 1.2.03, k=63) [v1.1] to assemble the genome. Assembly statistics were calculated with custom Perl scripts and the assembly graph was visualized in Cytoscape [v3.4.0].

Initial nanopore sequencing and data processing

We produced long reads from genomic DNA using Oxford Nanopore Technologies (ONT) R6 and R7 chemistry and aligned the long reads to the contigs exported by Velvet using LAST (version 4.60)[8]. We used simple settings: gap existence and extension penalties, mismatch penalty, and match reward all equal to 1. Additionally, we used the parameter -m 1000 to increase the alignment hit length until the hit occurs no more than a thousand times on the reference. This increases the number of alignments reported by sacrificing the precision, leading to a more reliable contig tiling across reads while allowing some erroneous alignments in the final report.

ONT sequencing

We have isolated additional genomic DNA from *Agrobacterium* strain LBA9072 strain using QIAGEN gravity-flow columns and produced another sequencing dataset with 400ng high molecular weight gDNA using R9.4 chemistry. We used a Rapid kit library preparation (SQK-RBK004) according to the manufacturer's protocols (Oxford Nanopore Technologies, Oxford, UK) that allows for swift preparation (approximately 10 minutes) and sequenced for 48 hours granting MinKNOW software (v19.06.8) control to the MinION sequencing device.

Assembly with Unicycler, Flye and Canu

Long-read data used for assembly were filtered on both length and quality.
The Canu assembly was performed with >1,000 bp reads without quality threshold, a minimum overlap of 500 bp and a 1,000 reads target coverage for read correction.
Unicycler and Flye have been restricted to use reads >3,000 bp that surpass the read quality threshold >10 PHRED, on a modest desktop (7 GB RAM and 8 CPU's) running Ubuntu 16.04 LTS. Additionally, we have provided a 5 Mbp genome size estimate. Flye (V2.4.2)[9] was used to perform *de novo* assembly using ONT data under default settings and using a minimal overlap length of 4,000 bp before considering merging contigs together. Unicycler (V0.4.7)[10] first uses Spades[11] to generate a short-read based assembly graph and performs error correction using short read data, then uses long reads to scaffold short-read based contigs.
Here assembly mode 'normal' (default) was used which is a setting that produces a balanced trade-off between genome completeness and assembly correctness. Gaps between contigs from high quality sequences are then filled in with long-reads, error corrected with Racon[12] and polished with Pilon (v1.18)[13]. Similar data filtering settings were provided, using only reads >3,000 bp with qualities >10 PHRED.

Transposon count and assembly similarity verification using Mauve

Mauve (v2.4.0)[14] was used to perform full genome progressive pairwise alignments to identify similarity among *de novo* assembly results, and between assemblies and reference strain *Agrobacterium vitis* S4. For circular assembly sequences we have reordered base positions and repositioned the cut generated by the individual assemblers to facilitate homology visualization. Additionally, we aligned two transposon sequences to the final Unicycler assembly result to identify the number of occurrences and location of those repeat sequences.

Insertion element identification with ISfinder software

ISfinder[15] was used to identify insertion sequences in the Unicycler assembly. ISfinder uses BLAST queries against a database of insertion elements to identify the family the insertion element originated from, as well as the covered length and homology identity to the reference. We have restricted identification to a minimum length of 300 bp, equivalent to a short protein of 100 amino acids and exceeding the length of Illumina reads.

Assembly visualizing

We used the Circos package (v0.69)[16] to visualize results in comparison to our Unicycler assembly results. We have visualized sequence similarities to genes originating from the *Agrobacterium vitis* S4 reference genome. Genes are aligned to the Unicycler contigs, both start and end positions from full and partial alignments are then converted to .bed file format and used as input for the Circos visualization. Similarly, we have generated bed files for locations of insertion elements identified by ISfinder and locations of two target transposon sequences. Sequencing data coverage of both Illumina and ONT sequencing technologies come from alignment files, and finally, we have compared the initial Velvet assembly to the Unicycler contigs. We have used bed files to report start and end position that we have retrieved from Minimap (V2.17-r954-dirty)[17] alignment files.

— Results

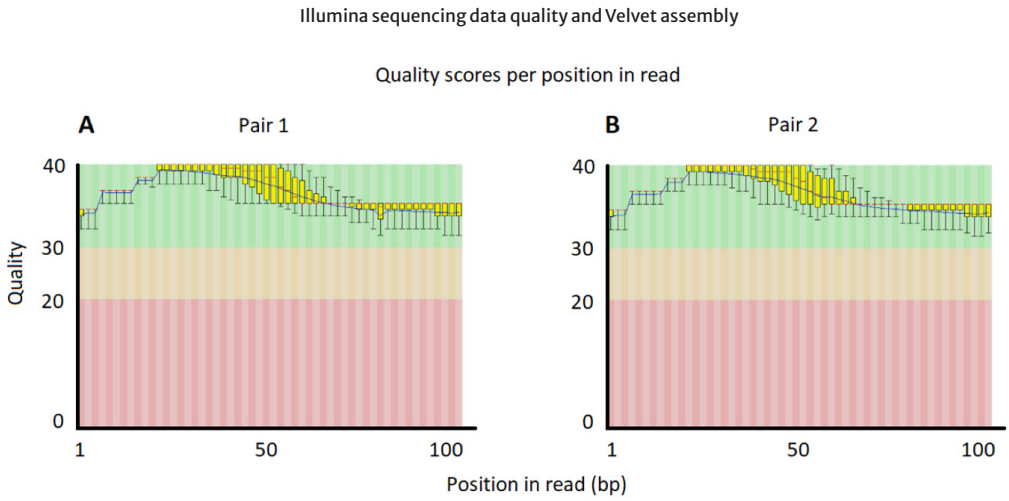


Figure 1 A–B Illumina sequencing data quality control of paired-end sequencing data.

Inspecting high-quality reads reveals sufficient quality for our paired-end dataset (>30 Phred). However, sequences are relatively short (max 100 bp in length), and quality drops are observed at the start and end of reads following the known sequencing characteristics corresponding to Illumina technology sequencing (Figure 1). Those reads are used as input sequences to generate a high-quality whole genome assembly using Velvet and combined with ONT sequencing data for hybrid assembly.

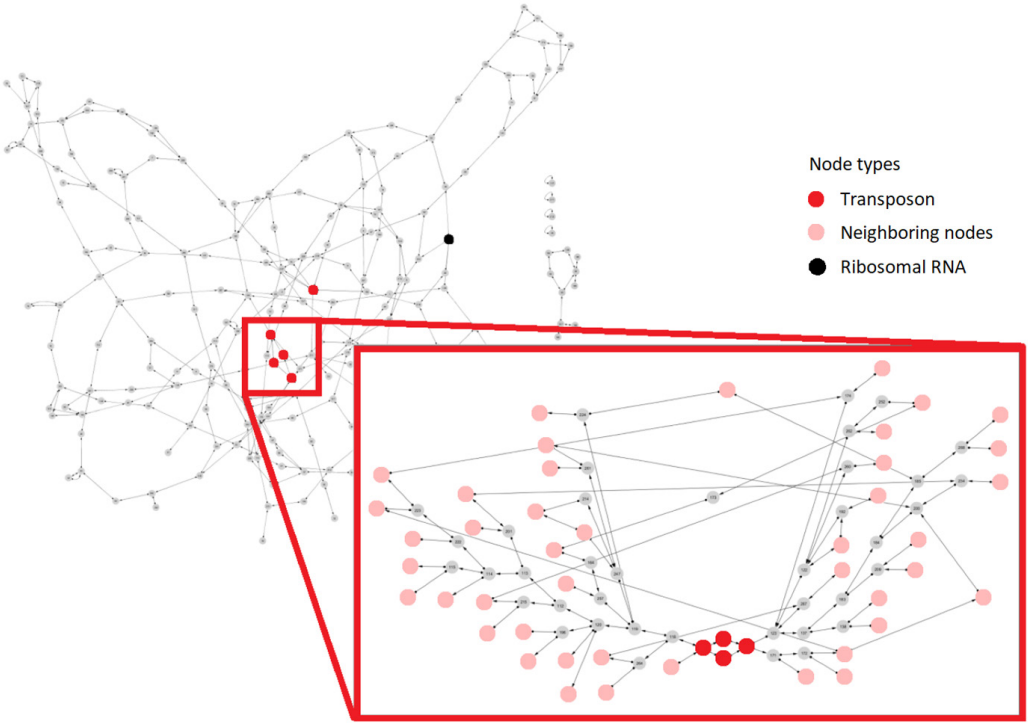


Figure 2 **High-quality whole genome assembly using Velvet.**
Complicated assembly graph due to present transposons indicated by red nodes. Ribosomal RNA contigs are depicted by a black node. The transposon depicted in the red box is split into four nodes due to single nucleotide differences. Contig nodes flanking the transposon are indicated in pink, grey indicates short (k-mer sized) nodes connected to neighbouring contigs.

Figure 2 shows contiguous sequences represented as nodes including the direction in which nodes are connected to their neighboring sequences (input and output directionality), and graph edges represent connections. Nodes that have increased coverage indicate repetitiveness and the coverage allows an estimation on how many times that sequence is observed throughout the complete assembly.

The Illumina-based assembly of strain LBA9072 strain is relatively fragmented, consisting of 154 contigs, with an N50 of 197,590 bp and a total assembly length of 5,872,508 bp (Velvet version 1.2.03, k=63, Table 3). Inspecting the Velvet assembly revealed the genome assembly graph is mainly complicated by the presence of two transposons, one of length 1,285 bp (Figure 2 zoomed-in section four red nodes) and another of length 935 bp (Figure 2 unboxed red node). These repeat contigs are connected to a multiplicity of neighboring nodes. The boxed repeat in red is connected to a total of 43 contigs, of which seven are connected on both left and right side of the repeat (Figure 2 zoomed-in section). Those 50 connections therefore initially suggest that this element is present 25 times throughout the genome. From here on it is referred to as the major transposon. The unboxed red repeat is similarly connected to a set of eight neighboring nodes and suggest the element is present eight times, from here on referred to as the minor transposon.

Sequences of nodes neighboring the major transposon (Figure 2 zoomed-in section pink nodes) are connected to tiny segments (Figure 2 zoomed-in section in gray), typically kmers representing 1-2 bp that do not result in contigs. A detailed view of neighboring nodes and transposon connectivity reveals the difficulty of resolving such context (Figure 2 zoomed-in section). Furthermore, the repeat in red itself is already split into 4 nodes, because of single nucleotide differences. In addition to this complexity, there are several copies of the 6.4 Kbp genes encoding ribosomal RNA (Figure 2 in black). The presence of this ensemble of repeat sequences results in the generation of a complex assembly graph, from which not a single complete plasmid or chromosome can be easily extracted using Illumina data alone.

Initial nanopore sequencing and data processing

From an experimental Oxford Nanopore Technologies sequencing run we obtained ~13x coverage, 13,158 sequencing reads, with a mean length of 5,611 bp and 12,211 bp N50 length. As the length of these reads often exceeds the lengths of the repeat elements, they could potentially be used to untangle complex contig connections that Illumina data alone cannot resolve (Figure 3 A)

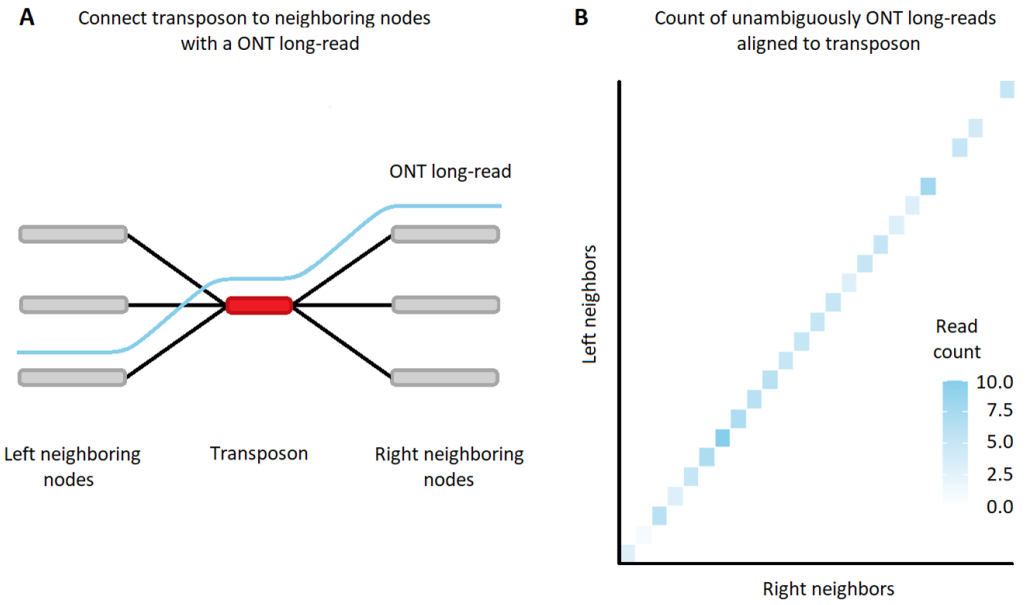


Figure 3 **A)** schematic representation of a long-read that spans over a repetitive element (**in red**) merging two previously unresolved regions (**in grey**) into a single large contig. **B)** read alignment count that connect left and right neighbors of the major transposon, alignment count ranges from one to ten reads and for two neighbors no alignment was observed.

Since a typical Velvet contig length is much larger compared to the read-length of this sequencing run most alignments are found in the middle of a Velvet contig. For repeat resolution we required a minimum of 2 independent reads that span over a repeat and connect the flanking contigs, in addition to an approximately correct distance between those contigs. Of the 13,158 reads, 483 aligned unambiguously to multiple contigs, and 585 links between contigs could be distilled. From those 585 links a subset is used to connect the major transposon: we analyzed 25 left and right neighboring nodes that are potentially connected to the major transposon (Figure 3 B). Rows represent the 'left' neighbors, columns the 'right'. For almost every neighbor, there exist sufficient unambiguous links (between 3 and 10 reads) to a single other neighbor. One link has low evidence (a single alignment), and for two neighbors no evidence for a connection was observed. In those latter cases, the final placement could not be resolved since neighboring nodes themselves were present twice in the genome.

Using this linkage information, we were able to give our assembly a significant upgrade. However, read-lengths from this sequencing run are insufficient to resolve the 6.4 Kbp ribosomal RNA repeat and these manual curations are very labor intensive. Hence using this low coverage ONT sequencing dataset a complete assembly remained unobtainable.

ONT sequencing

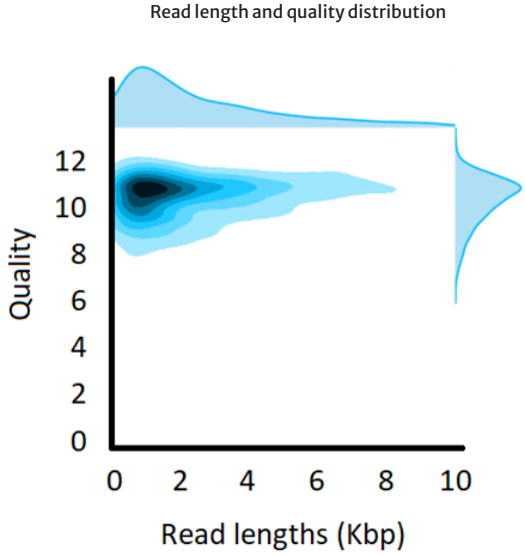


Figure 4 Data quality visualization; quality on Phred scale, where Phred 10 equals 10% error rate.

Table 1 ONT sequencing data statistics

Statistics	Count
Number of reads	170,955
Number of nucleotides (bp)	599,185,943
Maximum length (bp)	63,311
Mean length (bp)	3,504
Minimum length (bp)	1
Median length (bp)	2,121

We obtained 600 Mbp in approximately 171,000 reads of sequencing data, which corresponds to approximately 120x coverage of a 5 Mbp genome (Table 1). The average read-quality was better than 10 on the Phred scale (10% error or less) and mean read-lengths were around 3,500 bp (Figure 4). Read-length varies between 1 – 63,311 bp in length. These reads were filtered on quality and length and then used as input sequences for *de novo* assembly and combined with previously described Illumina sequencing data for hybrid assembly.

Two different assembly strategies

We have tested both long-read-only and hybrid *de novo* assembly strategies on this strain, with Canu and Fly using only ONT reads and Unicycler using both ONT and Illumina data. The resulting assemblies are similar in total genome size, number of contigs, contig lengths and sequence similarity. Using long-read data we have decreased the number of contigs to 4, 5 and 7 contigs for Flye, Unicycler and Canu assembly results, respectively, compared to the Velvet assembly counting 154 contigs (Table 2). Despite different contig number and N50 lengths, total assembly lengths remain similar between all four results. The N50 length is very comparable between Flye and Unicycler, but much lower for Canu. Upon closer inspection of assembled contigs (Figure 5), the main reason for this appears to be that Canu fails to assemble the large single chromosome into a circular contig, but instead outputs two linear contigs. In addition, Canu reports a set of smaller contigs (28,042 and 17,253 bp) that have no clear counterpart in either the Flye or the Unicycler results (Figure 6 B triple asterisk). Finally, since only Unicycler uses sequencing data from two distinct platforms, only Unicycler was able to reconstruct the 5,386 bp phiX174 Illumina spike-in viral genome.

Assembly visualization of two different assembly strategies

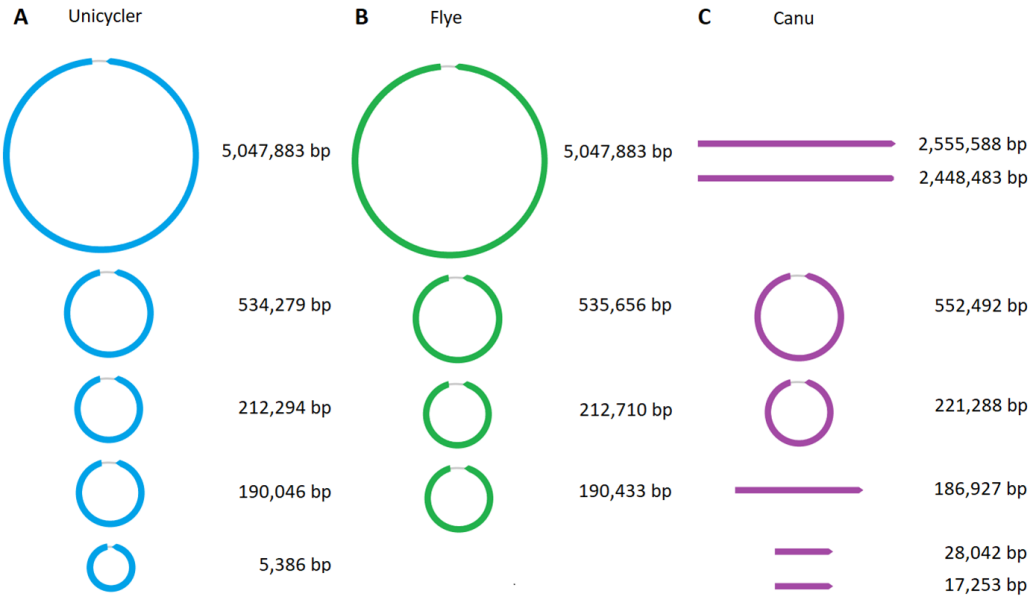


Figure 5 **A)** Unicycler assembly yields five circular contigs, one large chromosome and four additional plasmids. **B)** Flye assembly results in four circular contigs, one large chromosome and 3 additional plasmids similar in length compared to the Unicycler contigs. **C)** Canu outputs 7 contigs; Canu failed to circularize the large chromosome introducing at least two cuts that generate two individual linear contigs. Similarly, a 3 Kbp region is absent from the 190 Kbp plasmid, hence it remains linear.

Table 2 **Assembly statistics**

Statistics	Velvet	Unicycler	Flye	Canu
Number of contigs	154	5	4	7
Total assembly length	5,872,508	5,989,889	5,999,624	6,030,074
Contig N50	197,590	5,047,884	5,060,825	2,468,483

Verifying assembly sequence similarity using Mauve and counting transposon copies
We subsequently investigated the structural quality of the assembled contigs by assessing their similarity to an *Allorhizobium* vitis strain with a fully assembled genome (refered to as S4)[18]. Comparing the assembly results to the S4 reference strain shows overall high sequence and structural similarity between all three assembly results. A striking feature is that the assembled chromosomes have high similarity to the two chromosomes of the reference, and that the largest differences are observed across the plasmids. This suggests that most essential genes are conserved on the chromosome and that the bacterium harbors 'accessory' genes using plasmid sequences. Despite the well-conserved structure between assembly and reference some genomic rearrangements are present. Due to those rearrangements and two additional cuts generated by the Canu assembler the alignment becomes rather complicated to interpret. A single large circular chromosome is presented for both Unicycler and Flye assembly results, whereas Canu results two linear contigs. Some genomic rearrangements are confirmed between all three assembly results. Among others, around locus 1.6 Mbp on our reference A. vitis S4 we observe a small region (Figure 6 A single asterisk) that is reversed and joined on the Unicycler, Flye and Canu assemblies (reverse sequences are depicted on the bottom side of the grey horizontal axis). On the Canu assembly this region is found at 4.7 Mbp since the two largest contigs are linear, hence sequence regions cannot be repositioned to facilitate the visualization (Figure 6 A). Another rearrangement is observed around locus 3.8 Mbp on the reference (Figure 6 A double asterisk) and corresponds to the same green region around 3.0 Mbp on the Unicycler and Flye assembly, and around 3.7 Mbp on the Canu assembly.

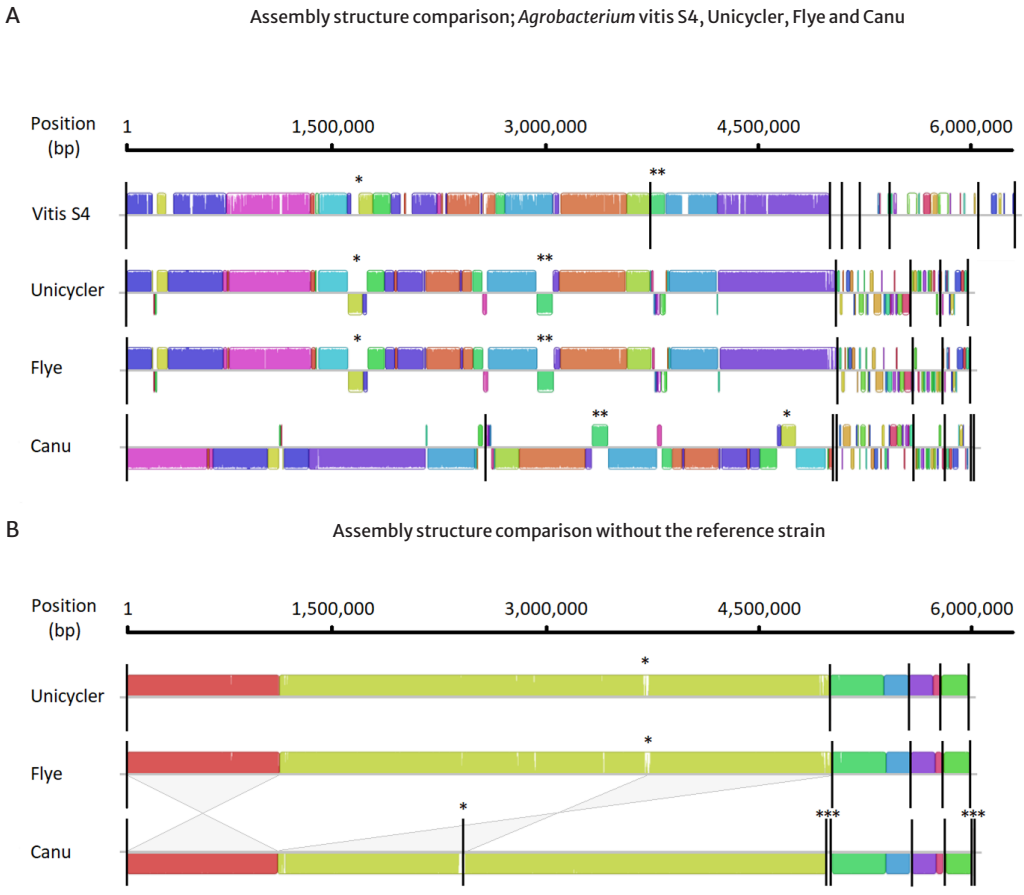


Figure 6 **A)** comparing the reference strain A. vitis S4 to all three assembly results. Coloring indicates similarity between segments, block heights indicate similarity between sequences. Black lines reveal contig ends and the grey horizontal line indicated directionality relative to the reference (forward orientation above the grey axis and reverse on the bottom half). **B)** The same assembly comparison without the S4 reference sequence to facilitate visualization.

Once we remove the reference strain and reorder the circular chromosomes from Unicycler and Flye, the visualization provides a much clearer impression on the quality of the assembly results. All three assemblies show high similarity to each other (Figure 6 colored block heights indicate sequence similarity). The two large linear Canu contigs are particularly misleading since a typical *Allorhizobium* genome comprises two larger chromosomes. Around 3.7 Mbp there exists a small region with low sequence identity between Flye, Unicycler and Canu assembly results (Figure 6 B asterisk), interestingly this is located on one of the cuts introduced by Canu. Finally, the Canu assembly is structurally similar to both Flye and Unicycler results, however, Canu outputs its final contigs in reverse order. The start and end regions of Flye and Unicycler contigs are reversed and merged on the Canu assembly (Figure 6 B in red and green shaded areas, respectively). Furthermore, a low similarity region observed at the same position for Unicycler and Flye assemblies, and around the boundary of the first Canu contig (Figure 6 B single Asterix) in addition to some very small contigs (Figure 6 B triple Asterix).

From the Velvet assembly we retrieved the sequences of the two most prominent repetitive regions. Sequencing data coverage initially suggested those regions were present 25 and 8 times throughout the whole genome. By aligning those sequences to the Unicycler contigs we were able to verify that there are 25 and 9 copies of those regions respectively (Figure 7), consistent with the initial estimates based on the Velvet graph (Figure 2).

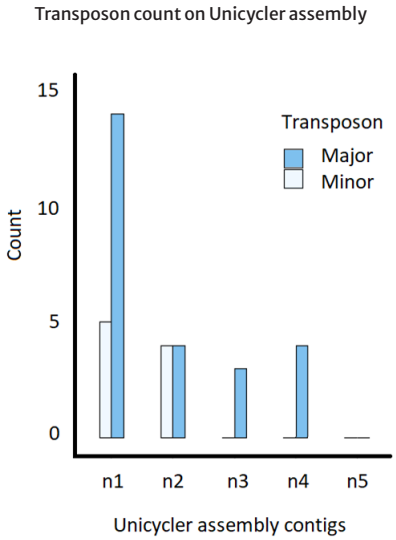


Figure 7 **transposon count on Unicycler assembly contigs.** The two transposon sequences retrieved from the Velvet assembly aligned to the Unicycler contigs. We found 25 copies of the major transposon and 9 copies of the minor transposon.

Canu discontinuities based on ambiguous alignments

When aligning the two longest Canu contigs to the longest Unicycler contig and zooming in to the regions where the Canu contigs are disjoint we find either a small gap in between or a small overlap of the two Canu contigs. The top track indicates the locations of the Canu cuts on the Unicycler contig, followed by the position track in Mbp, the alignment position of gap and overlap region of Canu contigs and ambiguous read alignments on those regions (Figure 8 blue for ONT data and red for Illumina data). We found ambiguous alignments either in proximity (Figure 8 A) or directly on the cut location (Figure 8 B), for the gap and overlap region, respectively.
This means reads align to the visualized location but also elsewhere in the genome, highlighting an unresolvable decision for assemblers based only on these reads. Despite a generous 80 times coverage of unambiguously aligned reads on those locations (data not shown) Canu is unable to merge the two contigs. A potential explanation for breaking up a contig could be the presence of insertion elements. Those elements have a repetitive nature and could cause reads to align ambiguously and in turn complicate decision-making processes that eventually lead to the introduction of a cut. Insertion elements, among which the major transposon, are situated in proximity to the two loci where a cut is introduced (Figure 9 around 1,330 and 3,800 Mbp).

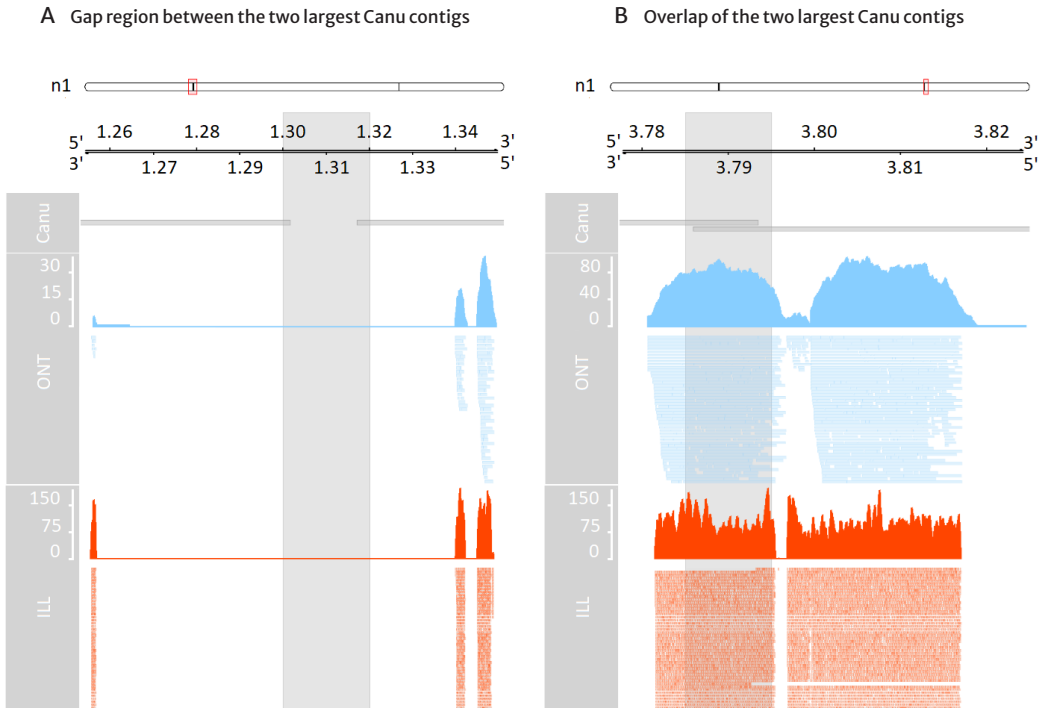
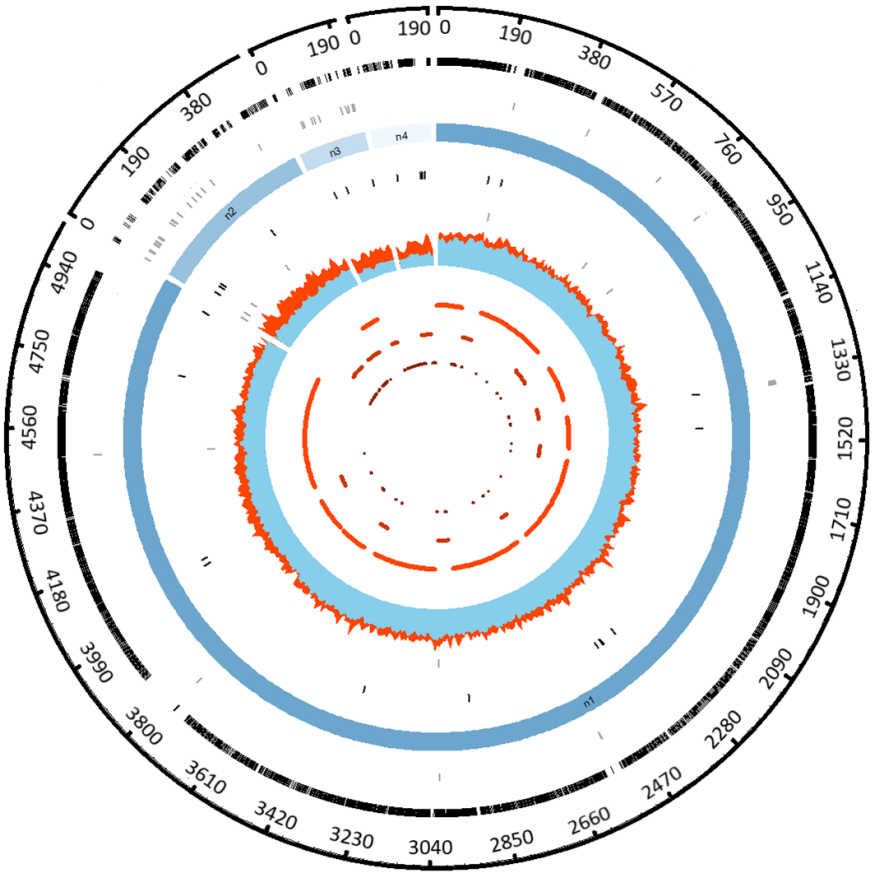


Figure 8 **A)** Two largest linear Canu contigs aligned to the largest circular Unicyler contig indicated at the top (n1). A small gap in between the two linear Canu contigs is present around 1.3 Mbp. Ambiguous alignments are represented in blue and red for ONT and illumina sequencing data, respectively. Some ambiguities are observed in proximity to the gap region.
B) Around 3.79 Mbp an overlapping region is present near the Canu cut, ambiguous alignments from both ONT and Illumina data, across and around the overlap, offer an explanation why Canu struggles to link those two contigs into a single sequence.

Unicycler assembly visualization and Velvet assembly comparison

Here we depict an overview of the Unicycler assembly contigs compared to our previous Velvet assembly results. From outwards to inwards (Figure 9), the position track in Kbp, followed by gene homology to the reference strain S4 (black). From the 5,433 genes provided from the reference strain we were able to align 4,400 to the Unicycler assembly. 3,993 genes on the n1 contig, 216 on the n2 contig, 91 on the n3 contig and 99 on the n4 contig. The largest contigs has a higher gene homology compared to the plasmid sequences, interestingly the surroundings of a few insertion sequences show a clear lack of gene similarity (e.g., around 190 Kbp, 400 Kbp, 2,500 Kbp, 3,800 Kbp). Interestingly, three of those loci are identified as insertion elements that are not classified as either the major or minor transposon. The absent gene similarity around those locations suggests that genomic diversification is facilitated by, and therefore found around, repetitive regions. Most insertion elements are observed inside the smaller contigs n2, n3 and n4.



Track positions

1	Position (Kbp)	2	Gene Homology (*A. vitis* S4)	3	ISfinder hits	4	Unicycler contigs	5	Major transposon
6	Minor transposon	7	Coverage ONT/ILL	8	>100 Kbp Velvet contigs	9	40 - 100 Kbp Velvet contigs	10	<40 Kbp Velvet contigs

Figure 9

Track 1; Base position (in Kbp).
Track 2; Shows the Unicycler homology to genes that originate from our reference strain A. vitis.
Track 3; Locations of insertion elements identified by ISfinder on the Unicycler contigs.
Track 4; Unicycler contigs (excluding n5 that is too small to visualize).
Track 5; Locations of the major transposon on the Unicycler contigs.
Track 6; Locations of the minor transposon on the Unicycler contigs.
Track 7; sequencing data coverage (ONT data in blue overlapping Illumina data in red).
Track 8; Velvet contigs >100 Kbp.
Track 9; Velvet contigs between 40 and 100 Kbp.
Track 10; Velvet contigs smaller then 40 Kbp.

The third track shows insertion sequences as found by the tool ISfinder. ISfinder identifies several different insertion sequences, the major and minor transposon are classified as insertion sequences from the IS5 family. However, it only reports a subset of the major and minor transposon locations compared to aligning the major and minor transposon sequences to the Unicycler contigs using Minimap2 (track five and six). Interestingly, BLAST results of the two transposon sequences from the Velvet assembly originate from the Rhizobium sp. 21/90 tumor inducing plasmid found in a Himalayan blackberry from Oregon USA at locus 104,603 - 105,694 and 188,695 – 189,528[19]. Those hits have >98% identity over the complete transposon region and are both annotated as IS5 family transposases. Alignment to the genes from the S4 reference did not result in significant hits, unless the similarity threshold was relaxed considerably (only 82% of the 935 bp was covered with <75% identity). The fourth track indicates the Unicycler contigs ordered from large to small (indicated by a blue shade from dark to light). The fifth and the sixth track reveal the locations of the major (black) and minor (grey) transposon copies that are aligned to the Unicycler contigs. Track seven is an overlay visualization of Illumina (red) and ONT (blue) data coverage. Interestingly, a large difference in coverage between ONT and Illumina data is observed for plasmid sequences, even though we have not performed a read-length selection. Finally, the eighth to tenth track depict the Velvet assembly results ordered from large to small contigs. Contigs >100 Kbp are indicated in bright red, followed by contigs between 100 Kbp and 40 Kbp and finally in dark red contigs <40 Kbp. The >100 Kbp track indicates the large chromosome is nearly complete, only a few gaps are introduced based on Illumina data alone. However, due to the repetitive structure some noise remains observed in the final track (small contigs overlap with the larger ones). High fragmentation is primarily observed for the plasmids, were some contigs range between 100 and 40 Kbp and many <40 Kbp contigs are observed. This represents low sequence complexity making it more difficult to assemble plasmids accurately.

References

1. Kuzmanović, N., Biondi, E., Overmann, J. et al. Genomic analysis provides novel insights into diversification and taxonomy of *Allorhizobium* vitis (i.e. *Agrobacterium* vitis). BMC Genomics 23, 462 (2022). https://doi.org/10.1186/s12864-022-08662-x

2. Anu Kalia, Nanotechnology in Bioengineering: Transmogrifying Plant Biotechnology, Omics Technologies and Bio-Engineering Volume 2: Towards Improving Quality of Life 2018, Pages 211-229 | https://doi.org/10.1016/B978-0-12-815870-8.00012-7

3. Indra A. Padikasan et al., Agricultural Biotechnology: Engineering Plants for Improved Productivity and Quality | https://doi.org/10.1016/B978-0-12-815870-8.00006-1

4. Eugene W. Nester, *Agrobacterium*: nature's genetic engineer, Front. Plant Sci., 06 January 2015 | https://doi.org/10.3389/fpls.2014.00730

5. Gustavo A. de la Riva, *Agrobacterium* tumefaciens: a natural tool for plant transformation, Electron. J. Biotechnol. v.1 n.3 Valparaíso dic. 1998 | http://dx.doi.org/10.4067/S0717-34581998000300002

6. J. S. Robalino-Espinosa, Segregation of four *Agrobacterium* tumefaciens replicons during polar growth: PopZ and PodJ control segregation of essential replicons, PNAS October 20, 2020 | https://doi.org/10.1073/pnas.2014371117

7. Jochen Gohlke et. Al., Plant responses to *Agrobacterium* tumefaciens and crown gall development, Front. Plant Sci., 23 April 2014 | https://doi.org/10.3389/fpls.2014.00155

8. Szymon M. Kiełbasa, et al., Adaptive seeds tame genomic sequence comparison | Published in Advance January 5, 2011, doi:10.1101/gr.113985.110Genome Res. 2011. 21: 487-49

9. Kolmogorov M., Yuan J., Lin Y. and Pevzner PA. (2019). Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology, 37(5), 540-546.

10. Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13(6): e1005595. https://doi.org/10.1371/journal.pcbi.1005595

11. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77. pmid:22506599

12. Robert Vaser et al., Fast and accurate *de novo* genome assembly from long uncorrected reads | Genome Res. 2017 May; 27(5): 737–746. doi: 10.1101/gr.214270.116

13. Walker BJ, Abeel T, Shea T, et al.: Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement | PLoS One. 2014;9(11):e112963. 10.1371/journal.pone.0112963

14. Aaron C.E. Darling et al., Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements | Genome Res. 2004 Jul; 14(7): 1394–1403. doi: 10.1101/gr.2289704

15. Siguier P. et al. (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34: D32-D36 | doi: 10.1093/nar/gkj014.

16. Krzywinski, M. et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res (2009) 19:1639-1645

17. Heng Li, Minimap2: pairwise alignment for nucleotide sequences | Bioinformatics, Volume 34, Issue 18, 15 September 2018, Pages 3094–3100, https://doi.org/10.1093/bioinformatics/bty191

18. Steven C. Slater et al., Genome Sequences of Three *Agrobacterium* Biovars Help Elucidate the Evolution of Multichromosome Genomes in Bacteria | DOI: https://doi.org/10.1128/JB.01779-08

19. Alexandra J Weisberg et al., Diversification of plasmids in a genus of pathogenic and nitrogen-fixing bacteria | Philos Trans R Soc Lond B Biol Sci. 2022 Jan 17;377(1842):20200466. doi: 10.1098/rstb.2020.0466. Epub 2021 Nov 29.