BMC Microbiology

RESEARCH Open Access

A contamination-reduced genome of *Plasmodiophora brassicae* with comprehensive functional annotation

Afsaneh Sedaghatkish[1], Meik Kunz[2], Bruce D. Gossen[3] and Mary Ruth McDonald[1*]

Abstract

Plasmodiophora brassicae causes clubroot disease in brassica crops worldwide but cannot be cultured outside its host, complicating its genome sequencing. Previous efforts relied on infected plant tissues and subtracting host DNA, which left microbial contamination and reduced the accuracy of the genome sequence. Single-cell genome sequencing (SCS) offers a promising solution by isolating individual cells free of host of microbial DNA and improving assembly accuracy. We applied a single-cell genome sequencing approach to ~ 4,000 protoplasts isolated from a highly virulent *P. brassicae* pathotype 5X population collected in Canada. The *de novo* assembly yielded a 23.3 Mb genome containing 8,758 predicted genes. Functional annotation using InterProScan identified 176,410 total domain hits, representing 6,809 distinct domain annotations and covering nearly all predicted genes, a substantial improvement compared to earlier annotations that covered only 34–57% of genes. We also identified 22 genes associated with virulence-related domains. A comparison with the previous reference genome revealed the *de novo* assembly of single cells was ~ 8% smaller, suggesting the removal of 1.9 Mb of contaminant sequences. This study demonstrates the value of SCS for overcoming contamination challenges in obligate, soil-borne pathogens, and establishes a framework to apply similar approaches to other non-culturable microbes.

Keywords Single-cell sequencing, Obligate pathogens, Microbial contamination, Clubroot, Canola

Author summary

Clubroot is a serious disease that affects brassica crops such as canola and cabbage, causing major losses for farmers. The disease is caused by a parasite called *Plasmodiophora brassicae*, which lives inside plant roots. Studying the DNA of this parasite is very challenging because it cannot be grown outside the plant, and previous efforts to sequence its genome were often contaminated by the plant's own DNA and other microbes. In this study, we used an innovative technique called single-cell genome sequencing to isolate and analyze the DNA from thousands of individual parasite cells, minimizing contamination. This approach allowed us to produce a much cleaner and more complete genome sequence. We identified nearly 9,000 genes, including many that may play a role in how the parasite causes disease. Our findings demonstrate that single-cell sequencing can overcome obstacles in studying hard-to-grow pathogens and could help scientists better understand similar organisms in the future.

*Correspondence:
Mary Ruth McDonald
mrmcdona@uoguelph.ca
[1]Department of Plant Agriculture, University of Guelph, Guelph, ON N1G 2W1, Canada
[2]The Bioinformatics CRO, Inc., Sanford, FL 32771, United States
[3]Agriculture and Agri-Food Canada, Saskatoon Research and Development Centre, Saskatoon, SK S7N 0X2, Canada

Introduction

Microorganisms play vital roles in ecosystems, agriculture, and human health. However, many remain challenging to study due to their inability to grow under laboratory conditions. One such organism is *Plasmodiophora brassicae* Woronin, a plant pathogen and obligate biotroph that completes nearly its entire life cycle within living host cells. Only the primary and secondary zoospore stages are motile in soil; all remaining developmental stages occur within host root tissues. This dependency on host-derived nutrients and signaling molecules makes *P. brassicae* unculturable in artificial media. Efforts to grow it outside its host have consistently failed, largely due to its complex life cycle, strict requirement for living plant tissue, and inability to reproduce independently of host roots. These biological constraints explain the limitations of traditional culture-based methods and underscore the need for alternative approaches, such as single-cell sequencing [1–3]. Standard approaches may not provide accurate sequencing of the genomes of such non-culturable organisms, particularly of obligate plant pathogens with complex life cycles and complex cell wall structures. These challenges are compounded by the difficulty in isolating high-quality DNA free from contamination by host or environmental microbes in soil-borne organisms. As a result, the genomes of many important plant pathogens remain poorly characterized, which limits the understanding of their biology and potentially the development of effective management strategies.

Plasmodiophora brassicae is an obligate chromist soil-borne pathogen that causes clubroot, an important constraint in the production of brassica crops worldwide, including economically important species such as canola (*Brassica napus* L.) [3]. *Plasmodiophora brassicae* has a complex life cycle with two main phases, primary and secondary infection (Fig. 1) [2].

The cycle begins when dormant resting spores in the soil germinate under favorable conditions, releasing motile zoospores that infect root hairs, initiating the primary infection phase. Inside root hairs, the pathogen forms large multinucleate cells called plasmodia. The plasmodia differentiate to produce secondary zoospores

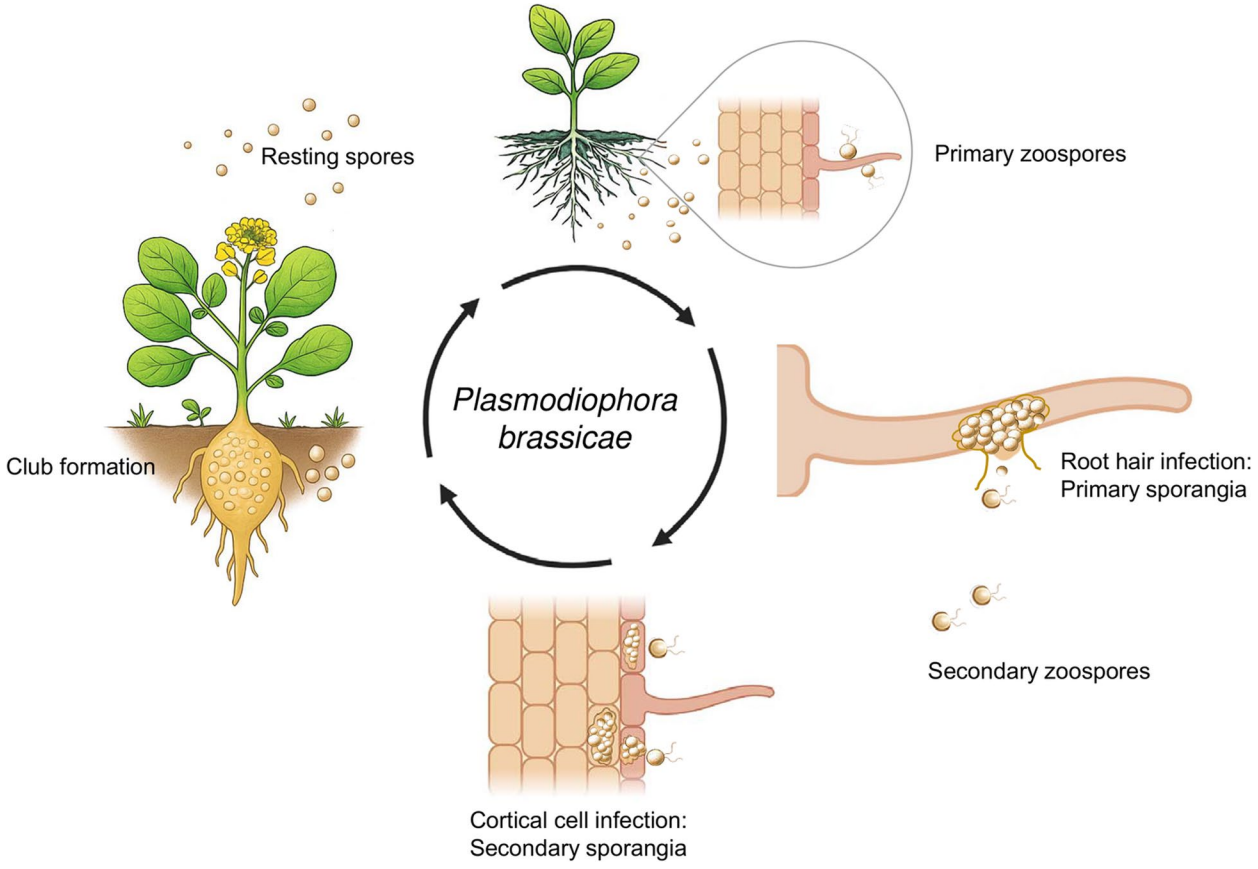

Fig. 1 *Plasmodiophora brassicae* life cycle. Resting spores germinate in soil under favorable conditions, releasing motile primary zoospores that infect root hairs (primary infection phase). Inside root hairs, the pathogen forms multinucleate primary plasmodia, which then differentiate to release secondary zoospores. These secondary zoospores infect cortical cells, forming secondary plasmodia that induce hypertrophy and hyperplasia, leading to the formation of characteristic root clubs. Within these clubs, the pathogen develops new resting spores that are released into the soil after plant decay, completing the cycle. This figure is originally created for this study, adapted from the descriptions in Kageyama and Asano [2] and Siemens et al. [3]

that are released from the root hairs to infect cortical tissues, which is the secondary infection phase [3, 4]. Colonization of the root cortex by secondary plasmodia after secondary infection results in excessive cell proliferation and formation of clubroot symptoms. The secondary plasmodia later mature and form resting spores. When the plant dies the plant tissues decay and the new resting spores are released into the soil, completing the cycle and ensuring the long-term survival of the pathogen. A single infected root (club) can harbor millions to billions of resting spores, contributing to the spread of the disease [3]. Clubroot management currently relies on the deployment of resistant cultivars, but these are often based on single-gene sources of resistance that are vulnerable to breakdown. Over time, new virulent pathotypes of *P. brassicae* have emerged, capable of overcoming this resistance, often within just a few growing seasons [3, 5].

The rapid emergence of new, virulent pathotypes indicates substantial genetic diversity within *P. brassicae* populations. However, the genetic basis of virulence remains poorly understood, largely due to limited genomic knowledge and a lack of insight into the complex molecular interactions between the pathogen and host resistance [1, 6]. Reliable molecular markers are needed for pathotype identification, but so far only a few pathotypes can be separated using markers [7–10]. The genetic analysis of *P. brassicae* has been hindered by several challenges inherent to obligate pathogens. These pathogens complete their life cycle within host cells, which prevents researchers from isolating the pathogen in pure culture for analysis. Also, DNA of this soil-borne pathogen, extracted from infected plant tissues, is often contaminated with both host DNA and environmental soil microorganisms, making it difficult to obtain a clean genome sequence. Previous genome sequencing efforts have typically relied on bulked resting spores collected from clubbed roots, and removal of host DNA based on published reference genomes of the host, but these methods have been limited in their ability to remove the contamination from soil microorganisms [11–16].

One promising approach for overcoming these limitations is single-cell sequencing (SCS), a cutting-edge technology that enables the isolation and sequencing of individual cells. The SCS has been predominantly used in mammalian systems to study cancer, immunology, and developmental biology [17, 18], and its application to non-human organisms is rare, with only a few studies exploring its use for microorganisms [19, 20]. In the context of *P. brassicae*, SCS offers a powerful tool to bypass the contamination issues associated with bulk DNA extraction from infected roots. By isolating individual protoplasts derived from resting spores, this method enables the sequencing of distinct genotypes within a single infected root, providing a more accurate picture of the pathogen genome. Additionally, SCS can help generate high-quality reference genomes that are less prone to contamination from host and soil microbes [21, 22].

In this study, we used the output from single-cell sequencing to produce a *de novo* assembly of the *P. brassicae* genome. By developing a tailored bioinformatics pipeline for the *de novo* assembly, we were able to generate an improved (cleaner and therefore more accurate) reference genome for use in molecular studies of *P. brassicae*. This research marks a significant advancement in the genomics of *P. brassicae*. In addition, the methods established in this study have broader implications for genomic research on other non-culturable, soil-borne pathogens, offering new opportunities for studying the genome of previously recalcitrant microorganisms.

Results

Quality assessment and contiguity of the *de novo* assembly

The quality of the *de novo* assembly was assessed by comparing key metrics, including contiguity and GC content distribution, to the most recent reference genome for *P. brassicae*, GCA_036867785.1_ULAVAL_Pb3A_ genomic (subsequently known as the reference genome in this paper). Assembly performance was evaluated by comparing our assembly with the ULAVAL_Pb3A reference (Fig. 2A), analyzing contig length distribution (Fig. 2B), and examining GC content patterns (Fig. 2C).

The *de novo* assembly generated a raw total length of 45.7 Mbp, which is slightly shorter than the reference genome (47.8 Mbp). The N50 value for the assembly was 1.205 Mbp, indicating high-quality genome assembly. The assembly contained 345 contigs (assembled sequences), slightly more fragmented than the reference genome, which has 310 contigs. Overall, the resulting genome assembly covers 92% of the reference genome, with even distribution across all chromosomes. The total assembly size is 23.3 Mb.

The contiguity of the *de novo* assembly was evaluated through cumulative contig length distribution (Fig. 2B). The cumulative length curve for the *de novo* assembly (red line) closely followed but consistently fell below the reference genome (blue line), across the length distribution. This indicates a smaller maximum contig size and higher fragmentation relative to the reference genome.

When the GC content distribution of the *de novo* assembly was compared to the reference genome (Fig. 2C), both genomes showed a similar GC content profile, with peaks centered around 58–60% GC content. *The de novo* assembly exhibited a slightly narrower distribution and a lower number of windows at the peak compared to the reference genome. These results indicated that the base composition of the *de novo* assembly was largely consistent with the reference genome, with minimal GC bias.

A

Assembly	*De novo* assembly	GCA_036867785.1 ULAVAL_Pb3A_genomic
# contigs (>= 0 bp)	20	20
# contigs (>= 1000 bp)	20	20
# contigs (>= 5000 bp)	20	20
# contigs (>= 10000 bp)	20	20
# contigs (>= 25000 bp)	20	20
# contigs (>= 50000 bp)	20	20
Total length (>= 0 bp)	23372281	25283981
Total length (>= 1000 bp)	23372281	25283981
Total length (>= 5000 bp)	23372281	25283981
Total length (>= 10000 bp)	23372281	25283981
Total length (>= 25000 bp)	23372281	25283981
Total length (>= 50000 bp)	23372281	25283981
# contigs	20	20
Largest contig	1995923	2102640
Total length	23372281	25283981
GC (%)	59.56	59.37
N50	1205706	1371253
N90	693889	747440
auN	1305345.1	1396604.3
L50	8	8
L90	17	17
# N's per 100 kbp	0	0

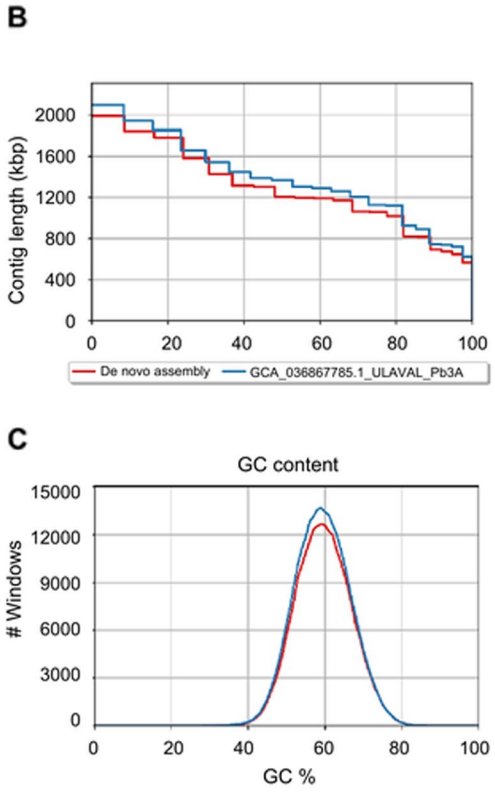

Fig. 2 *De novo* assembly quality assessment. **A** Assembly statistics comparing the IDBA-UD assembly with the ULAVAL_Pb3A reference**. B** Contig length distribution showing assembly contiguity and (**C**), GC content distribution across the assembled genome. Contig-level statistics are provided in Supplementary File S1

Chromosome coverage and unmapped reads in *de novo* assembly

The average coverage of chromosomes in the newly assembled genome was consistently high across all chromosomes, ranging between 85 and 97% (Fig. 3A and S1 file). Minor differences were noted at CP145415.1 and CP145418.1 on individual chromosomes, which exhibit slightly lower coverage (~ 85%). Overall, the uniformity of coverage indicated a high-quality assembly process that efficiently captured most regions across the genome.

The total base counts for each chromosome in the *de novo* assembly (red bars) closely mirrored the reference genome (blue bars) for most chromosomes, indicating a high degree of completeness (Fig. 3B). However, slight differences were apparent in a few chromosomes, such as CP145411.1, CP145417.1, and CP145421.1, where the *de novo* assembly has fewer bases than the reference genome. Notably, larger chromosomes like CP145409.1 and CP145410.1 show a minor reduction in base counts in the *de novo* assembly but remained within an acceptable range for comparative genomic analyses. In comparative genomics, variations of up to ± 10% in chromosome size are generally considered acceptable, and all differences observed here fall within this threshold.

To investigate the annotation similarity between two genomes, the chromosomal alignments were visualized in the *de novo* assembly (B, orange) relative to the reference genome assembly (A, blue). Primary contigs were scaffolded by connecting the contigs into larger chunks using homology between mutant contigs and the reference genome. Both genomes exhibit broad alignment across all chromosomes, with the reference genome consistently showing larger portions of the chromosomes compared to the *de novo* assembly (Fig. 3C). Chromosome length is expressed in megabase pairs (Mbp), with the largest chromosome (CP145409.1) extending to ~ 2.0 Mbp. Grey shaded regions indicate areas of synteny (conserved blocks of genetic content). There was substantial conservation between the two assemblies; only slight variations are observed, particularly toward the chromosome ends where syntenic regions appear truncated or missing in the *de novo* assembly. Chromosomes CP145409.1 through CP145418.1 have near-complete alignment between the two genomes, with *de novo* assembly exhibiting minor gaps. However, white regions indicate contaminated regions in the ULAVAL reference genome with main regions on chromosomes CP145412.1, CP145413.1, CP145417.1, CP145429.1 and CP145427.1. Chromosomes CP145419.1 to CP145428.1

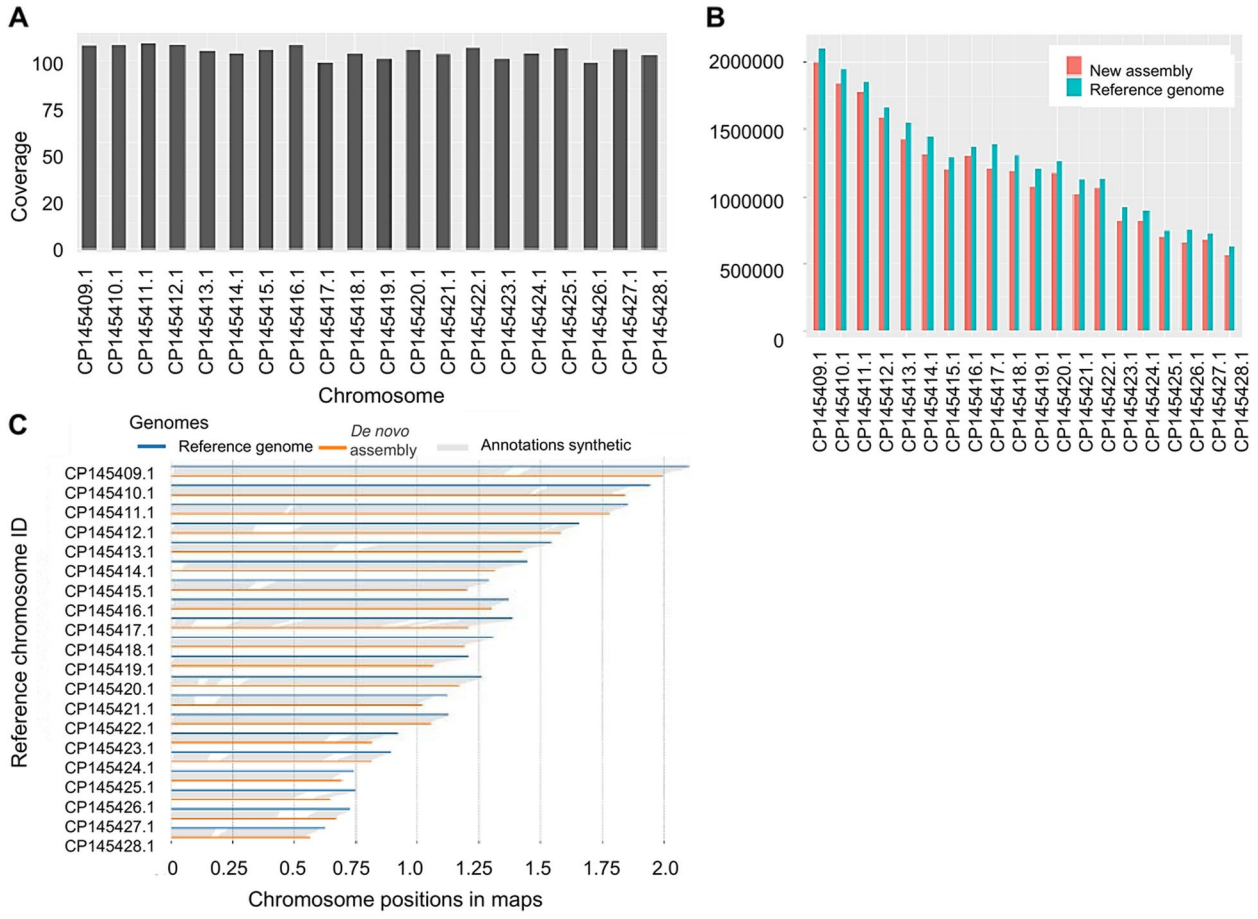

Fig. 3 Chromosome coverage in the *de novo* assembly. **A** genome coverage and (**B**) chromosome size similarity between the GCA_36867785.1_ULAVAL_Pb3A_genomic (blue) and *de novo* assembly (red). The y-axis represents the total number of aligned bases for each chromosome. **C** similarity in chromosome position between the GCA_36867785.1_ULAVAL_Pb3A_genomic (blue) and *de novo* assembly (orange). The horizontal axis indicates relative chromosome positions (map units) ranging from 0 to 2.0

display more pronounced differences, as *de novo* assembly shows reduced alignment lengths compared to the reference genome. Notably, CP145423.1 and CP145428.1 are among the shortest chromosomes, with *de novo* assembly aligning to only ~ 60% and 50% of the reference genome length, respectively. Variations in annotation coverage suggest that these represent potential regions of divergence between the two genomes.

Genome-wide distribution and density of annotated features

We further ran gene annotation using Braker which identified 8,758 unique genes in the *de novo* assembly of *P. brassicae* (Figs. 4, 5). The distribution of genomic features across multiple regions (P145409.1 to P145428.1) revealed variable densities of annotated elements (Fig. 3). Each region, represented as a horizontal bar, spans lengths ranging from 0.5–1.5 Mbp. The green vertical bars indicate regions enriched with annotated genomic features, while black bars represent feature-poor or unannotated regions. Regions such as P145409.1 and P145410.1 exhibited a higher density of green bars, suggesting a uniform distribution of annotated features across the segments. In contrast, regions such as P145426.1 and P145428.1 displayed larger proportions of black bars, indicative of low annotation density or gaps in feature detection. Notably, longer genomic regions, such as P145409.1 and P145418.1, displayed a dispersed yet consistent distribution of features. Shorter segments, such as P145424.1 and P145426.1, contained more pronounced clusters of features interspersed with larger feature-poor regions. Overall, the presence of localized clusters of genomic features was observed across most segments, with sporadic black regions occurring more frequently toward the distal portions of the shorter genomic regions. These findings highlight variability in annotation completeness or genomic complexity, warranting further investigation to determine whether these regions correspond to coding sequences, non-coding elements, or repetitive genomic regions.

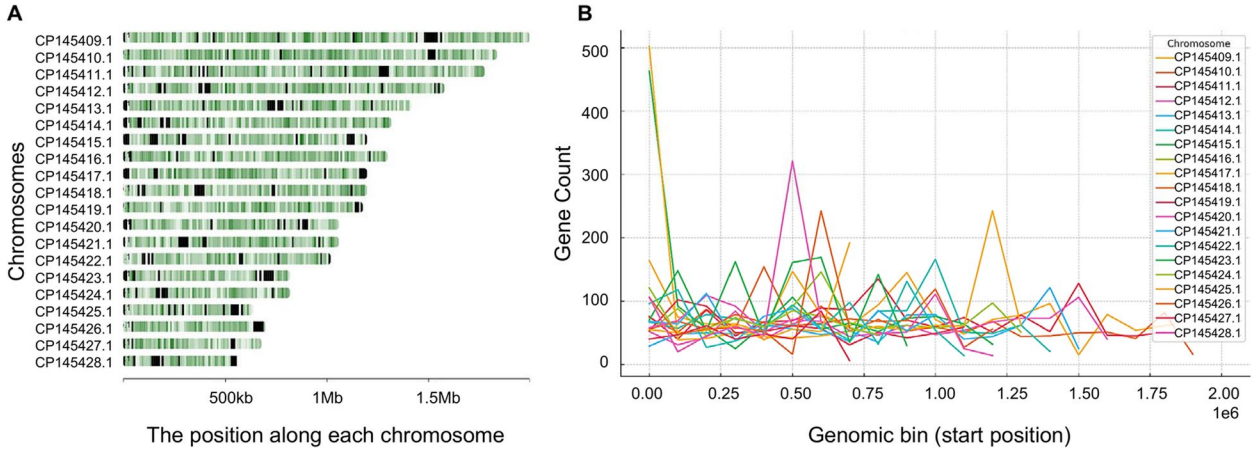

Fig. 4 Contigs and telomeres in *Plasmodiophora brassicae* P5X strain. **A** Density of coding sequences (CDS, shown in green) and gaps (shown in black) across the genome, calculated in non-overlapping sliding windows of 10 kilobases (kb). The x-axis shows genomic position (bp), and the y-axis indicates feature density. The intensity of the color is proportional to the density of coding or gap sequences. **B** Gene density across chromosomes, calculated in non-overlapping 100 kb windows. The x-axis represents the genomic bin start position along each chromosome, and the y-axis shows the number of predicted genes in each window. Each colored line corresponds to a different chromosome (CP145409.1–CP145428.1)

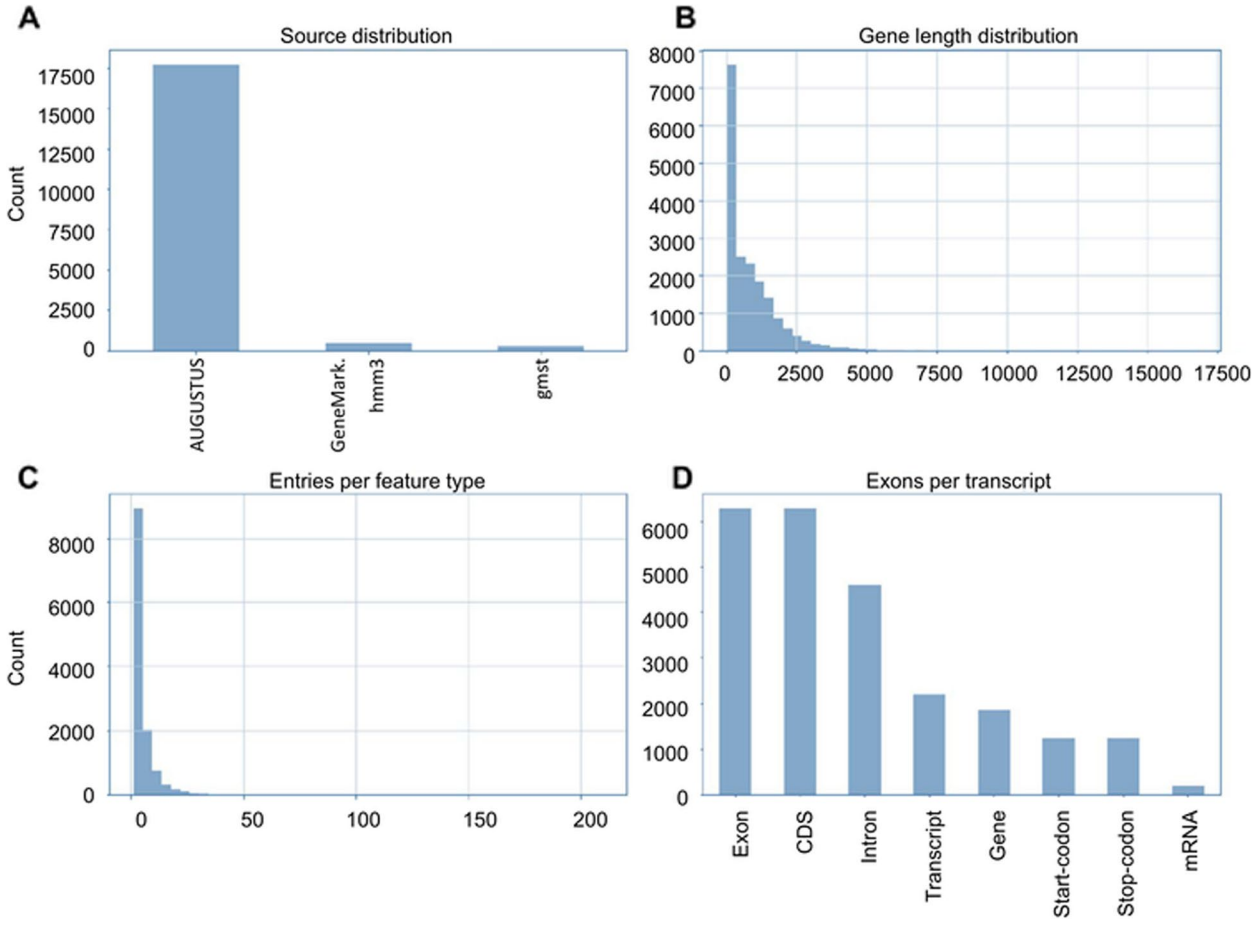

Fig. 5 Distribution of the identified 8,758 genes in the *de novo* assembly of *P. brassicae*. **A** Distribution of gene predictions across different annotation tools, with AUGUSTUS contributing the majority of entries. **B** Histogram of gene length showing a right-skewed distribution, with most genes under 2,000 bp. The horizontal axis indicates gene length (bp). **C** Frequency of annotation feature entries per gene, highlighting variation in gene structure complexity. The horizontal axis represents feature length (bp). **D** Count of different annotation feature types, showing high representation of exons and CDSs, typical of eukaryotic gene models

Genome completeness was further assessed using BUSCO analysis (compleasm v0.2.7). Our assembly recovered 80.62% complete BUSCOs, with 6.2% fragmented and 13.18% missing, indicating comparable completeness to other recently published P. brassicae genomes.

To further quantify structural variation, gene density was calculated across all chromosomes using 100 kb non-overlapping windows (Fig. 4B). This analysis revealed gene-rich intervals interspersed with gene-sparse regions, reflecting the heterogeneous architecture of the *P. brassicae* genome. Chromosomes such as P145409.1 and P145410.1 showed relatively consistent gene densities, while P145426.1 and P145428.1 exhibited patchier profiles with extended intervals lacking annotated genes. These feature-poor zones may reflect repetitive or unassembled regions, potentially linked to centromeric or subtelomeric loci. Gene-dense regions appeared concentrated in central chromosomal areas and declined near telomeric ends. This spatial distribution further supports the observation of localized complexity and suggests underlying differences in chromatin organization or gene function across the genome.

Gene structure and annotation metrics in the *de novo* assembly

A total of 8,758 genes were identified from the *de novo* assembled genome, with the majority of predictions generated using the AUGUSTUS annotation pipeline [23] (Fig. 5A and S2 file). Gene lengths displayed a right-skewed distribution, with most genes measuring under 2,000 bp and a few extending beyond 10,000 bp (Fig. 5B) and the corresponding protein length distribution followed a similar pattern (S3 file). The number of annotation entries per gene varied widely, with most genes having fewer than 20 features, although a subset exhibited more than 200 (Fig. 5C), indicating diversity in gene structure. Feature-type analysis revealed that exons and coding sequences (CDSs) were the most common annotations, followed by introns, transcripts, and genes (Fig. 5D). The presence of start- and stop-codons and mRNA entries further supports the accuracy and completeness of the annotation process.

To evaluate proteome complexity, we analyzed the length distribution of predicted protein sequences. The resulting histogram showed a right-skewed profile, with most proteins falling between 100 and 800 amino acids in length (S3 file). A small number of proteins exceeded 1,000 amino acids, suggesting the presence of large multi-domain proteins, which may be associated with structural or regulatory roles in the pathogen.

Functional domain annotation

We performed functional annotation of the *Plasmodiophora brassicae* genome using InterProScan, incorporating multiple curated domain databases including Pfam, Gene3D, SMART, SUPERFAMILY, PANTHER, and MobiDBLite for disorder prediction (S4 and 5 files) [24]. This analysis identified 6,809 functional annotations (unique GeneID-distinct annotaions) across the predicted gene set, capturing diverse biological roles associated with metabolism, mitochondrial processes, fungal adhesion, and virulence.

In total, InterProScan generated 176,410 domain hits (redundant domain calls), reflecting multiple domain assignments per gene, and identified 12,309 distinct protein domain types, indicating extensive domain diversity across the proteome (S5 file). The most frequent annotations originated from MobiDBLite (114,578 entries), followed by Gene3D (11,898), Pfam (11,150), and SUPERFAMILY (9,507).

Importantly, 22 genes were identified that were associated with putative virulence or pathogenicity functions, based on keyword-matched domain descriptions. Notable examples included the pathogenesis-related protein signature, T6SS phospholipase effector Tle1-like domain, Spermadhesin CUB domain, and Aflatoxin B1 aldehyde reductase, suggesting effector-like mechanisms that may play roles in host–pathogen interactions. These features highlight candidates previously underrepresented in public annotations of *P. brassicae.* BLASTP alignment of these genes against known effector databases revealed similarity to previously characterized virulence-associated proteins, including homologs of ToxB, Nep1-like proteins, and phospholipase effectors. Most of these domains are conserved across multiple *P. brassicae* pathotypes, though some appear enriched in pathotype 5X, suggesting potential diversification of effector repertoires in highly virulent strains.

Comparison with prior annotations reveals enhanced functional coverage

Compared to earlier annotations of *P. brassicae* genomes, such as the reference isolate e3 described by Schwelm et al. [15], Bi et al. [25], Rolfe et al., [14], and GCA_003833335.1 (NCBI RefSeq annotation), our dataset demonstrates both substantially improved completeness and functional resolution of the annotation. Previous studies reported approximately 3,000–5,000 genes with assigned functions and fewer than 5,000 total domain hits, primarily due to the use of limited or single-database pipelines. In contrast, our annotation approach, leveraging eleven integrated InterProScan databases, enabled the characterization of all 8,758 predicted genes with both high-resolution domain information and expanded functional breadth (Table 1).

Table 1 Comparison of functional genome annotation metrics between the current study and previously published reference genomes

Feature	Current Study	Previous Estimates	References
Genome assembly size	23.3 Mb	25.1–26 Mb	[13–15]
Predicted genes	8,758	8,800–9,000	[13, 15, 25]
Functionally annotated Genes	8,758 (100%)	~3,000–5,000 genes with GO/Pfam/KOG terms	[14, 15, 25], NCBI RefSeq (GCA_003833335.1)
Distinct annotations	6, 809 (176,410 raw hits) via InterProScan annotations	< 5,000	NCBI GenBank; Ensembl Fungi; UniProt
Distinct domain types	12,309	< 5,000	[25] Ensembl Fungi
Unique gene function pairs	32,169	~3,000–5,000	[14, 15, 25], NCBI RefSeq (GCA_003833335.1)
Annotation tools used	InterProScan (Pfam, Gene3D, SMART, etc.)	NCBI PGAP, UniProt, GOA (partial)	GOA; InterPro databases
Uncharacterized genes	52 (0.5%)	4000–5300 (45–60%)	[15, 25] NCBI
Virulence/pathogenicity domains	22 domain-based hits only	50–100 predicted effectors (via SignalP, EffectorP)	[15, 25, 26]

Importantly, all functionally annotated genes in the *de novo* assembly carried at least one informative domain assignment, suggesting a high level of functional resolution across the genome. Only 52 genes (0.5%) lacked detectable functional domains in our dataset, compared to 45–60% uncharacterized gene content in earlier studies. These improvements provide a more comprehensive foundation for exploring pathogenicity, gene regulation, and host–microbe interactions in *P. brassicae*.

Our mapping results indicated that ~ 8% of the total reads did not align with the reference genome, suggesting the presence of non-target sequences. Taxonomic classification of these unmapped reads revealed that they predominantly belonged to soil microbes, comprising a diverse set of mostly fungal taxa. The majority of these reads mapped to fungal species (12,828 hits), with a smaller proportion assigned to bacteria (1,191 hits) and an extremely low presence of viral sequences (2 hits) (Fig. 6A).

Further analysis of the most frequently detected microbial taxa (Fig. 6B) identified *Cryptococcus gattii* WM276 as the most abundant species (1,050 hits), followed by.

Eremothecium gossypii (744 hits), and *Colletotrichum higginsianum* (661 hits). Several other fungal pathogens commonly found in soil were also detected, including *Phaeosphaeria nodorum* (wheat septoria blotch), *Melampsora laricis-populina* (rust fungus), *Fusarium graminearum* (Fusarium head blight and *Fusarium oxysporum* (wilts and rots on many plant species)), and *Botrytis cinerea* (grey mold). The presence of these species indicates that there was contamination from the soil microbiome or the clubbed root sample. These microbes were found in the 8% of non-mapping reads. Fragments between 100–1,000 bp typically allow reliable genus-level assignment in BLAST, and fragments > 500 bp generally support species-level classification, providing confidence in the taxonomic composition identified.

Methods

Protoplast isolation

A single root exhibiting clubroot symptoms was selected from a highly virulent *Plasmodiophora brassicae* collection originally isolated from canola cultivated at the AAFC Research Farm in Normandin, Quebec (approximately 48.83° N, 72.53° W) [27]. This isolate was propagated from a single club on a susceptible cultivar of Shanghai pak choi (*Brassica rapa* var. *chinensis*) and subsequently bulked through multiple cycles. The pathotype of the isolate was determined in this study using the Williams differential host set [28], which identified it as predominantly pathotype 5X. The designation 'X' indicates the ability to overcome first-generation clubroot resistance in canola [16, 29].

Resting spores were extracted using a standard preparation method [30]. In brief, the infected root was rinsed under running tap water, surface-sterilized with 30% commercial bleach (Clorox®, 6% sodium hypochlorite) for 10 min, followed by immersion in 70% ethanol for 2 min., and then homogenized in sterile distilled water using a laboratory blender. The homogenate was passed through ten layers of cheesecloth and further purified by filtration through a 30 µm filter (Millipore Sigma, Oakville, Canada) to eliminate fine debris. Resting spores were counted using a hemocytometer. The suspension was centrifuged three times at 100 g for 1 min to remove residual supernatant, followed by a final centrifugation at 1000 g for 5 min at room temperature. The resulting pellet was resuspended in sterile deionized water [27].

To reduce potential external genomic DNA contamination, the spore suspension was treated with RQ1 DNase (Promega, ON) at 37 °C for 30 min, followed by enzyme inactivation at 65 °C using the manufacturer's stop solution.

Spore walls were digested to produce protoplasts by incubating the pellet in a lysing solution composed of Glucanex (lytic enzymes from *Trichoderma harzianum*, Sigma Aldrich, Canada) at 300 mg/mL in 0.7 M NaCl, sterile-filtered through a 0.2 µm syringe filter. The spore

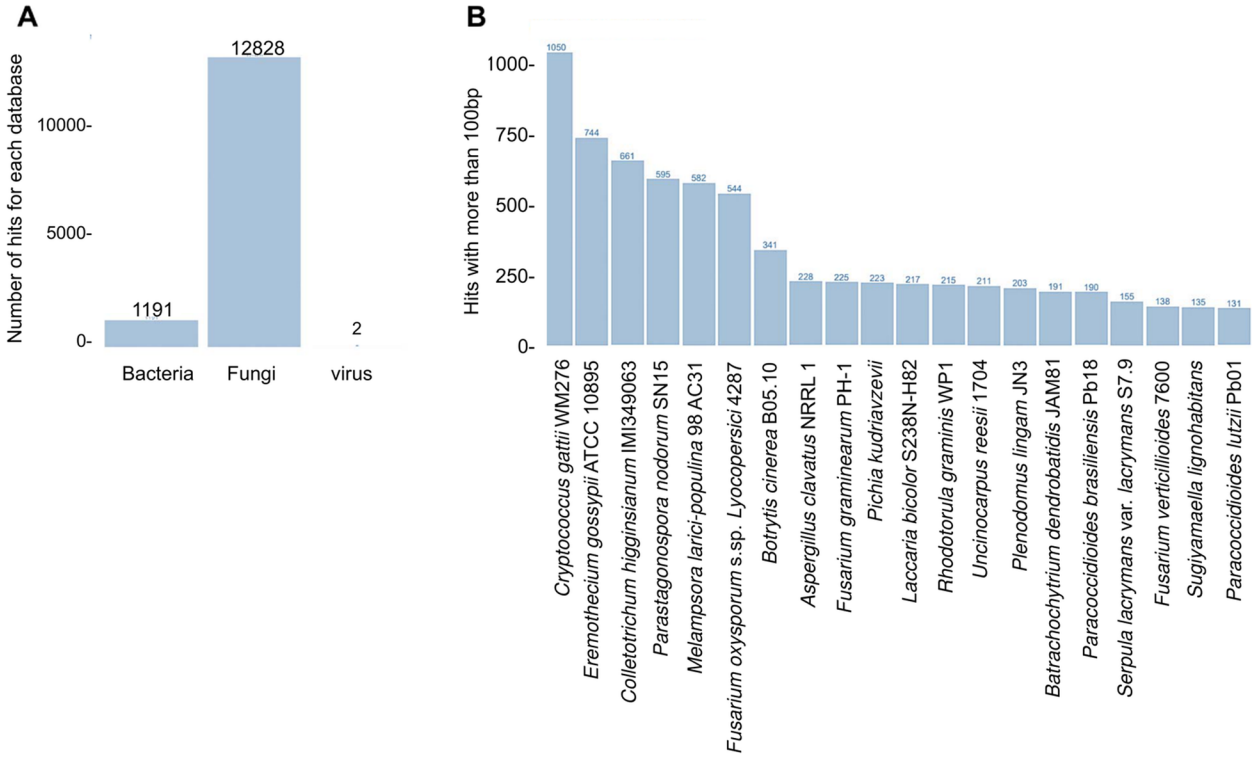

Fig. 6 NCBI nucleotide blast search for the 8% of non mapping reads to the reference genome. **A** The number of hits per bacteria, fungi and virus database. **B** top 20 species per number of hits

concentration was adjusted to 1×10^8 spores/mL. The mixture was incubated at 30 °C with gentle agitation (40 rpm) for 7 h. Protoplast development was monitored under a light microscope. Following digestion, the suspension was centrifuged at 3000 rpm for 5 min at 4 °C, and the supernatant was removed. The pellet was gently resuspended in 40 mL of ice-cold STC buffer (1.2 M sorbitol, 10 mM Tris pH 7.0, 50 mM $CaCl_2$), pH-adjusted to 7.5. This concentration and exposure time were previously determined to be optimal for protoplast release from resting spores [27].

The protoplasts were maintained on ice for the remainder of the protocol. Cell counts were obtained using a hemocytometer, and the suspension was diluted to 1×10^7 protoplasts/mL in STC buffer. The final preparation was flash-frozen in liquid nitrogen and transported on dry ice for further processing.

Single-cell DNA sequencing

Single-cell library preparation and sequencing were conducted by the McGill Genome Centre (Montreal, QC, Canada) as a service. Frozen protoplast aliquots were thawed on ice and centrifuged at 3000 × g for 5 min. The pellet was resuspended in buffer containing 1 M mannitol, 10 mM Tris (pH 7.0), and 50 mM $CaCl_2$. Cell viability and morphology were assessed using the LIVE/DEAD Viability/Cytotoxicity Kit (Thermo Fisher Scientific), and DNA content was evaluated using DRAQ5 staining (Thermo Fisher Scientific).

Approximately 4,000 protoplasts were used to generate a single-cell DNA library using the Chromium Controller and Single Cell DNA Reagent Kit (10 × Genomics), following the manufacturer's instructions (Protocol CG000153, Rev C) [27]. Library clean-up was performed using the SPRIselect Reagent Kit (Beckman Coulter), and quality control for fragment size and yield was assessed via LabChip GX (Perkin Elmer). Quantification was performed using the KAPA Library Quantification Kit for Illumina platforms. The library was sequenced on an Illumina NovaSeq SP flowcell using the following read structure: 100 bp Read 1, 8 bp Index 1 (i7), 0 bp Index 2 (i5), and 100 bp Read 2 [27]. We obtained ~ 24.7 million mapped bases across a ~ 16.0 Mb genome, resulting in an average sequencing depth of 1.54 ×. To further assess coverage distribution, all quality-trimmed reads were aligned to the reference genome, and read depth at each position was calculated with SAMtools. This analysis showed a true mean sequencing depth of 74 ×, with 65% of the genome covered at ≥ 50 × and 30% at ≥ 70 ×. Reads flagged as duplicates and low-complexity reads were removed to reduce amplification artefacts and improve assembly accuracy. Although the per-cell coverage is well within the expected range for single-cell genome sequencing (~ 1 × –2 ×) [31], the sequencing of ~ 4,000

individual protoplasts yielded a cumulative depth of 74×. This overall coverage was sufficient to support gene prediction, domain-level annotation, and comparative genomic analysis. Of the 3,957 barcoded protoplasts, 99% contained more than 1,000 reads mapping to *P. brassicae*, indicating that nearly all droplets captured pathogen DNA. All demultiplexed reads were included in the initial *de novo* assembly, and non-target contigs were removed afterward through contamination filtering.

Data processing

Processing and filtering of reads were conducted using SAMtools (with BCFtools v1.19 and HTSlib v1.19.1) [32]. The ULAVAL_Pb3A reference genome (GCA_036867785.1) was downloaded from NCBI and reindexed using SAMtools to ensure compatibility with downstream alignment tools. QUAST was run on both the *de novo* assembly and the reference genome to allow standardized comparison of assembly quality. Whole-genome alignments between the *de novo* assembly and the reference were generated using Minimap2, and RagTag used these alignments for reference-guided scaffolding and contig ordering. To identify and quantify contaminant sequences, we used BlobTools2 for taxon-annotated GC–coverage visualization and Kraken2 for k-mer–based taxonomic assignment. Contigs classified as non-Chromista or showing bacterial/fungal signatures were removed before scaffolding. This approach identified 1,542 contaminated contigs > 100 bp (totaling 1.9 Mb), which were excluded from the final assembly. Relative to the ULAVAL_Pb3A reference genome, this represents an ~ 8% reduction in genome size attributable to removal of contaminant regions. Contaminated regions present in the reference but absent from the *de novo* assembly were verified using BlobTools2 plots and read-depth anomalies. We applied a hybrid two-step strategy for genome reconstruction: first, *de novo* assembly with IDBA-UD [33], which generated contigs, followed by reference-guided scaffolding using RagTag v2.1.0 [34] against the GCA_036867785.1_ULAVAL_Pb3A reference genome. Assembly with IDBA-UD was performed using a k-mer size range of 20–100 bp (incremented by 20), a minimum contig length cutoff of 200 bp, and the default error-correction module to minimize sequencing artefacts. The quality of the genome assembly was assessed with Bioawk v1.0 and QUAST [35], while synteny between the *de novo* contigs and the reference genome was evaluated using Minimap2 v2.23 [36, 37] and SyRi v1.4 [38]. Gene prediction and annotation were carried out with BRAKER v3.0.8 [39], and visualization was performed using the ggplot2 R package (v3.3.5) (Fig. 7). In this study, 'contigs' refer to primary assembled sequences, 'scaffolds' to contigs ordered and oriented relative to the reference, and 'assembly' to the final genome as a whole.

Functional annotation

Gene prediction was conducted using BRAKER v3.0.8, integrating both RNA-seq data from *P. brassicae* infective stages and ab initio predictions. We selected BRAKER over alternatives such as EGAPx and funannotate because it is well established for combining transcriptomic evidence with ab initio methods and has been shown to outperform MAKER2 and funannotate for fungal and protist pathogens. We performed functional annotation of the predicted protein-coding genes in the *Plasmodiophora brassicae* genome using InterProScan v5.59–91.0. Protein sequences derived from the gene models were analyzed against multiple domain databases, including Pfam, Gene3D, SMART, SUPERFAMILY, PANTHER, ProSiteProfiles, PRINTS, Coils, and MobiDBLite. The analysis was conducted using default parameters, and domain matches with valid e-values and complete feature annotations were retained. InterProScan outputs were parsed to identify domain descriptions, associated InterPro entries, Gene Ontology (GO) terms, and pathway annotations where available. We did not perform repeat masking prior to annotation because the current RefSeq assembly of *P. brassicae* has already been repeat-masked, allowing direct comparative analyses. Given the limitations of short-read scDNA sequencing in reconstructing repetitive elements, additional repeat masking was unnecessary for the objectives of this study.

To ensure genome-wide coverage, all 8,758 predicted genes were included in the analysis, yielding 176,410 total domain assignments and 6,809 unique functional annotation events after filtering for redundancy. Functional domains associated with key biological processes, including metabolism, mitochondrial activity, adhesion, and virulence, were manually reviewed. A subset of virulence- and effector-related genes was identified based on keyword-matching of domain descriptions (e.g., "effector", "virulence", "toxin") and curated for relevance to host–pathogen interactions. Functional annotations were integrated with gene models (GTF) to enable downstream mapping and visualization.

Discussion

The *de novo* assembly exhibited a high degree of similarity to the reference genome in terms of GC content and overall contiguity. While the total assembly size and N50 values were slightly lower than those of the reference genome, the consistency in GC distribution and cumulative contig lengths indicated a reliable assembly with good quality, suitable for downstream analyses. In addition, the reliability of the *de novo* genome assembly was confirmed with high coverage and minimal discrepancies

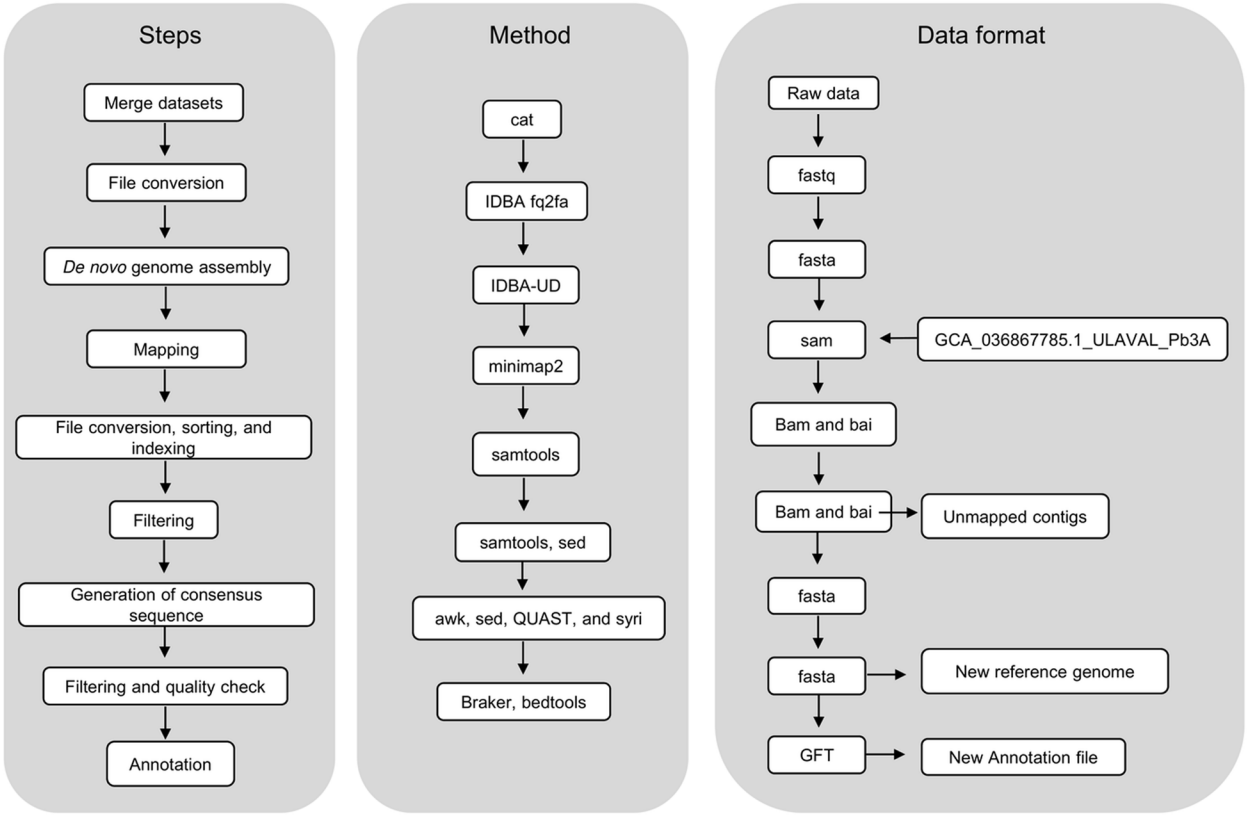

Fig. 7 Overview of the genome data processing pipeline. The diagram illustrates the sequential steps (left column), tools and methods used (middle column), and corresponding data formats (right column) for de novo genome assembly, mapping, consensus sequence generation, and genome annotation

in base counts compared to the reference genome. These findings underscore the robustness of the assembly process, particularly for larger, more complex chromosomes. Compared with previous Illumina/PacBio bulk-DNA assemblies (Table 1), the single-cell sequencing approach reduced contaminant content by ~ 8% while maintaining comparable contiguity, demonstrating a clear methodological improvement when sequencing obligate pathogens. Single-cell sequencing is prone to amplification bias, including allelic dropout, uneven coverage, and formation of chimeric sequences. To mitigate these issues, we applied stringent filtering to remove duplicate and low-complexity reads and optimized k-mer assembly parameters. While residual bias cannot be fully excluded, these steps helped ensure assembly robustness, particularly in repetitive regions. Assembly completeness was further supported by BUSCO analysis, which identified 80.6% complete single-copy orthologs. While slightly lower than other *P. brassicae* assemblies such as Pb3A, these results confirm that the single-cell sequencing approach produced a robust genome assembly suitable for downstream analyses. Although BUSCO completeness (80.6%) is slightly lower than in long-read assemblies, the functional annotation depth is substantially higher because InterProScan integrates eleven domain databases and captures functional information even for genes lacking orthologous BUSCO markers. Thus, BUSCO scores and annotation richness reflect different aspects of genome quality.

The comparative analysis demonstrates that conserved syntenic regions are observed across most chromosomes except that *de novo* assembly produced slightly shorter chromosomes. The detection of unmapped reads in the *P. brassicae* single-cell dataset highlights the challenge of microbial contamination in host–pathogen sequencing studies of soil-borne obligate pathogens.

In addition to structural genome analysis, comprehensive functional annotation revealed an expanded view of *P. brassicae*'s proteome. InterProScan-based domain analysis assigned 6,809 non-redundant functional annotations across all 8,758 predicted genes, supported by more than 176,000 total domain hits and 12,309 distinct protein domain types. These results substantially exceed the annotation coverage of previous studies, which reported functions for only 3,000–5,000 genes. The use of integrated domain resources such as Pfam, Gene3D, and SMART uncovered extensive protein diversity linked to metabolism, mitochondrial activity, and cellular binding functions. Within this diversity, 22 genes contained virulence- or effector-related domains, including

pathogenesis-related protein signatures, T6SS phospholipase effectors, and spermadhesin CUB domains, pointing to underexplored candidate mechanisms of host manipulation. These virulence-associated genes highlight conserved effector families such as ToxB and NLPs, suggesting that pathotype 5X retains core effectors while also diversifying specific loci to overcome host resistance. Together, these findings reinforce the value of high-resolution annotation in obligate pathogens and provide a framework for downstream functional validation and comparative pathogenicity studies.

Extent and limitations of functional grouping

Although all 8,758 predicted genes received at least one InterProScan-derived annotation, not all domain assignments are equally informative. In contrast to previous annotations in *P. brassicae*, where 45–60% of genes remained uncharacterized [15, 25], our annotation reduced this to 0.5% of the gene set, reflecting a significant improvement in functional resolution.

The comprehensive nature of our functional annotation, incorporating eleven domain databases, allowed the assignment of biologically meaningful functions to the vast majority of genes. Nevertheless, the presence of uncharacterized entries points to the continued challenge of annotating pathogen-specific or lineage-restricted proteins, which often lack homologs in curated domain repositories. These uncharacterized genes may represent novel effectors or virulence factors and should be prioritized in future studies involving expression profiling, structure prediction, or host-interaction assays. In contrast to the repeat-masked RefSeq assembly, our short-read assembly likely underrepresents highly repetitive elements. However, broad repeat content trends were consistent between assemblies, suggesting that major differences in coding regions are not due to repeat masking artefacts.

The 8% of reads that did not align to the reference genome were identified as microbial sequences, primarily from fungi and some bacteria commonly associated with rhizospheres. Among the most abundant taxa identified as contaminants were soil- and plant associated fungi, including *Cryptococcus gattii*, *Botrytis cinerea*, and *Fusarium oxysporum*, all of which are known to persist in plant environments [40–43]. These findings indicate that, despite careful sample preparation, the isolation of *P. brassicae* from clubbed roots may also co-purify DNA from soil microbes. This microbial background could stem from residual rhizosphere communities, endophytic associations, or contaminants from the symptomatic (clubbed) tissue itself. The contamination-reduced claim is based on the use of single-cell sequencing, in which each protoplast is isolated and barcoded individually, minimizing the risk of co-purified host DNA. Of the ~ 8% of reads not mapping to the reference genome, BLAST analysis revealed that most corresponded to soil fungi (e.g., *Cryptococcus gattii, Fusarium oxysporum*), consistent with residual rhizosphere DNA rather than contamination of the assembled genome itself. While contamination is an inherent challenge in sequencing from environmental samples, the current study highlights the usefulness of single cell sequencing for genomic studies. The presence of abundant fungal species in the unmapped reads also suggests that clubbed roots produced as the result of colonization by *P. brassicae* may open a niche for other soil-borne pathogens, raising questions about potential microbial interactions in clubroot-infected root tissues. In addition to host and soil microbial DNA, some of the contaminating sequences detected in earlier *P. brassicae* assemblies, and in our preliminary datasets, could possibly have originated from fungal endophytes in root tissues [44] that coexist with the pathogen. These fungi can persist within root tissues without causing disease and may contribute to the complex microbial environment of clubroot galls. The single-cell sequencing approach used in this study minimized such contamination, providing a cleaner and more accurate assembly of the *P. brassicae* genome [44]. Future studies could explore the functional roles of co-isolated microbes in clubroot symptom development and breakdown and dispersal of resting spores from clubs to determine whether these contaminants represent opportunistic pathogens, symbionts, or passive soil inhabitants.

Conclusion

Our findings establish a high-confidence genomic foundation for *Plasmodiophora brassicae*, with improved accuracy and functional grouping. The annotated genome offers new opportunities to investigate the molecular basis of virulence and host interactions in this pathogen. Beyond *P. brassicae*, the approach demonstrated here provides a valuable framework for genomic studies of other pathogens, especially obligate, non-culturable organisms facing similar technical challenges. Similar single-cell or single-organism sequencing approaches are also applicable to other obligate, non-culturable plant pathogens, including *Plasmodiophora betae* (sugar beet clubroot pathogen), *Spongospora subterranea* (powdery scab of potato), *Olpidium brassicae* (a root-infecting virus vector), and rust fungi such as *Puccinia graminis*, *Puccinia striiformis*, and *Melampsora lini*. These pathogens share key biological constraints strict host dependence, biotrophic intracellular stages, and the inability to culture them outside their host making them strong candidates for single-cell genomics–based approaches.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12866-026-04716-2.

Supplementary Material 1

Acknowledgements
We thank Haig Hugo Vrej Djambazian, Ashot Harutyunyan, and Ioannis Ragoussis from McGill Genome Center for cell barcoding and single cell sequencing.

Authors' contributions
A.S designed the studies, performed the experiments, analysed and interpreted the results and wrote the paper. M.K performed the bioinformatics analysis and contributed to the interpretation of the result. B.D.G and M.R.M. contributed to the interpretation of the results, edited the manuscript and secured funding for the study.

Funding
The author(s) declare that financial support was received for the research and/or publication of this article. This study was supported by Agriculture Development Fund of the Government of Saskatchewan and SaskCanola. Agriculture Development Fund of the Government of Saskatchewan (AGR-17881) and SaskCanola (ADF# 20200155).

Data availability
The genome assembly have been deposited in NCBI under accession number SAMN53902068. The raw sequencing reads have been deposited in the Sequence Read Archive (SRA) under accession SRR29917097. In addition, the sequences and annotation files from this study have been made available on Zenodo (https://doi.org/https://doi.org/10.5281/zenodo.15593496).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 24 October 2025 / Accepted: 2 January 2026
Published online: 19 January 2026

References

1. Dixon GR. The occurrence and economic impact of *Plasmodiophora brassicae* and clubroot disease. J Plant Growth Regul. 2009;28(3):194–202. https://doi.org/10.1007/s00344-009-9090-y.
2. Kageyama K, Asano T. Life cycle of *Plasmodiophora brassicae*. J Plant Growth Regul. 2009;28(3):203–11. https://doi.org/10.1007/s00344-009-9101-z.
3. Siemens J, Bulman S, Rehn F, Sundelin T. Molecular biology of *Plasmodiophora brassicae*. J Plant Growth Regul. 2009;58(1):130–6. https://doi.org/10.1111/j.1365-3059.2008.01943.x.
4. McDonald MR, Strelkov SE, Gossen BD, et al. The role of primary and secondary infection in host response to *Plasmodiophora brassicae*. Phytopathology. 2014;104(9):1078–87. https://doi.org/10.1094/PHYTO-12-13-0332-R.
5. Strelkov SE, Hwang SF, Manolii VP, et al. Virulence and pathotype classification of *Plasmodiophora brassicae* populations collected from clubroot-resistant canola (Brassica napus) in Canada. Can J Plant Pathol. 2018;40(3):284–98. https://doi.org/10.1080/07060661.2018.1436615.
6. Diederichsen E, Frauen M, Linders A, Hatakeyama K, Hirai M. Status and perspectives of clubroot resistance breeding in crucifer crops. J Plant Growth Regul. 2009;28(3):265–80. https://doi.org/10.1007/s00344-009-9116-1.
7. Sedaghatkish A, Gossen BD, McDonald MR. Characterization of a virulence factor in *Plasmodiophora brassicae*, with molecular markers for identification. PLoS ONE. 2023;18:e0289842. https://doi.org/10.1371/journal.pone.0289842.
8. Yang H, Shen Y, Yu Y, Pan L, Zhang Z. Specific genes and sequence variation in pathotype 7 of the clubroot pathogen *Plasmodiophora brassicae*. Eur J Plant Pathol. 2020;157:239–50. https://doi.org/10.1007/s10658-020-01973-3.
9. Zhang H, Feng J, Hwang SF, Strelkov SE. Characterization of a gene identified in pathotype 5 of the clubroot pathogen *Plasmodiophora brassicae*. Phytopathology. 2015;105:787–93. https://doi.org/10.1094/PHYTO-10-14-0288-R.
10. Zheng J, Xu L, Xue S, Cao T, Zhang H, Feng J, et al. Characterization of five molecular markers for pathotype identification of the clubroot pathogen *Plasmodiophora brassicae*. Phytopathology. 2018;108:552–60. https://doi.org/10.1094/PHYTO-08-17-0278-R.
11. Fu H, Harding MW, Feng J, Strelkov SE. Most *Plasmodiophora brassicae* populations in single canola root galls from Alberta fields are mixtures of multiple strains. Plant Dis. 2020;104(1):116–20. https://doi.org/10.1094/PDIS-02-19-0290-RE.
12. Holtz MD, Hwang SF, Strelkov SE. Genotyping of *Plasmodiophora brassicae* reveals the presence of distinct populations. BMC Genomics. 2018;19:254. https://doi.org/10.1186/s12864-018-4639-3.
13. Javed MA, Mukhopadhyay S, Normandeau E, Brochu AS, Pérez-López E. Telomere-to-telomere genome assembly of the clubroot pathogen *Plasmodiophora brassicae*. Genome Biol Evol. 2024;16:evae122. https://doi.org/10.1093/gbe/evae122.
14. Rolfe SA, Strelkov SE, Links MG, et al. The compact genome of the plant pathogen *Plasmodiophora brassicae* is adapted to intracellular interactions with host *Brassica* spp. BMC Genomics. 2016;17:272. https://doi.org/10.1186/s12864-016-2597-2.
15. Schwelm A, Berney C, Dixelius C, et al. The *Plasmodiophora brassicae* genome reveals insights into its life cycle and ancestry of chitin synthases. Sci Rep. 2015;5:11153. https://doi.org/10.1038/srep11153.
16. Sedaghatkish A, Gossen BD, Yu F, Torkamaneh D, McDonald MR. Whole-genome DNA similarity and population structure of *Plasmodiophora brassicae* strains from Canada. BMC Genomics. 2019;20:744. https://doi.org/10.1186/s12864-019-6118-y.
17. Hou Y, Song L, Zhu P, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012;148(5):873–85. https://doi.org/10.1016/j.cell.2012.01.042.
18. Wang J, Song Y. Single-cell sequencing: a distinct new field. Clin Transl Med. 2017;6:1–11. https://doi.org/10.1186/s40169-017-0151-9.
19. Heywood JL, Sieracki ME, Bellows WK, et al. Capturing diversity of marine heterotrophic protists: one cell at a time. ISME J. 2011;5(4):674–84. https://doi.org/10.1038/ismej.2010.166.
20. Liu Z, Hu S, Campbell V, et al. Single-cell transcriptomics of small microbial eukaryotes: limitations and potential. ISME J. 2017;11:1282–5. https://doi.org/10.1038/ismej.2016.190.
21. Blainey PC. The future is now: single-cell genomics of bacteria and archaea. FEMS Microbiol Rev. 2013;37(3):407–27. https://doi.org/10.1111/1574-6976.12015.
22. Xu Y, Zhao F. Single-cell metagenomics: challenges and applications. Protein Cell. 2018;9(6):501–10. https://doi.org/10.1007/s13238-018-0544-5.
23. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–44. https://doi.org/10.1093/bioinformatics/btn013.
24. Finn RD, Coggill P, Eberhardt RY, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279–85. https://doi.org/10.1093/nar/gkv1344.
25. Bi K, Chen T, Zhang Y, et al. Comparative genomics reveals different expansion and evolution patterns of repeat sequences in *Plasmodiophora brassicae* genomes. Sci Rep. 2016;6:25300. https://doi.org/10.1038/srep25300.
26. Pérez-López E, Waldner M, Hossain M, et al. Effector candidates in *Plasmodiophora brassicae* and their role in clubroot disease. Virulence. 2020;11(1):730–47. https://doi.org/10.1080/21505594.2020.1768334.
27. Sedaghatkish M, Gossen BD, McDonald MR. The first single-cell sequencing of *Plasmodiophora brassicae* reveals genetic diversity and clonal dynamics. Front Microbiol. 2025;16:1581233. https://doi.org/10.3389/fmicb.2025.1581233.
28. Williams PH. A system for the determination of races of *Plasmodiophora brassicae* that infect cabbage and rutabaga. Phytopathology. 1996;56:624–6.
29. Sedaghatkish A. The genomic structure and management of *Plasmodiophora brassicae*. PhD thesis, University of Guelph. 2020. Available from: http://hdl.handle.net/10214/17833.

30. Sharma K, Gossen BD, McDonald MR. Effect of temperature on primary infection by *Plasmodiophora brassicae* and initiation of clubroot symptoms. Plant Pathol. 2011;60:830–8. https://doi.org/10.1111/j.1365-3059.2011.02458.x.
31. Woyke T, Doud DFR, Schulz F, et al. Decontamination of MDA reagents for single cell whole genome amplification. PLoS ONE. 2011;6(10):e26161. https://doi.org/10.1371/journal.pone.0026161.
32. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
33. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28(11):1420–8. https://doi.org/10.1093/bioinformatics/bts174.
34. Alonge M, Soyk S, Ramakrishnan S, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20:224. https://doi.org/10.1186/s13059-019-1829-6.
35. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5. https://doi.org/10.1093/bioinformatics/btt086.
36. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95. https://doi.org/10.1093/bioinformatics/btp698.
37. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.
38. Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20:277. https://doi.org/10.1186/s13059-019-1911-0.
39. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. Methods Mol Biol. 2019;1962:65–95. https://doi.org/10.1007/978-1-4939-9173-0_5.
40. Araújo AE, Lima GSA, Goulart LR, et al. Survival of *Botrytis cinerea* as mycelium in rose crop debris and as sclerotia in soil. Fitopatol Bras. 2005;30:516–21. https://doi.org/10.1590/S0100-41582005000500009.
41. Gordon TR. *Fusarium oxysporum* and the Fusarium wilt syndrome. Annu Rev Phytopathol. 2017;55:23–39. https://doi.org/10.1146/annurev-phyto-080615-095919.
42. McKeen CD, Wensley RN. Longevity of *Fusarium oxysporum* in soil tube culture. Science. 1961;134(3495):1528–9. https://doi.org/10.1126/science.134.3495.1528.
43. Tortora GJ, Funke BR, Case CL. Microbiology: An Introduction. 13th ed. Pearson. 2021.
44. Poveda J, Díaz-González S, Díaz-Urbano M, Velasco P, Sacristán S. Fungal endophytes of Brassicaceae: molecular interactions and crop benefits. Front Plant Sci. 2022;13:932288. https://doi.org/10.3389/fpls.2022.932288.
45 Zhou Q, Feng J, Hwang SF, Strelkov SE. A molecular marker for the specific detection of new pathotype 5-like strains of *Plasmodiophora brassicae* in canola. Plant Pathol. 2018;67:145–55. https://doi.org/10.1111/ppa.12724.

Publisher's Note