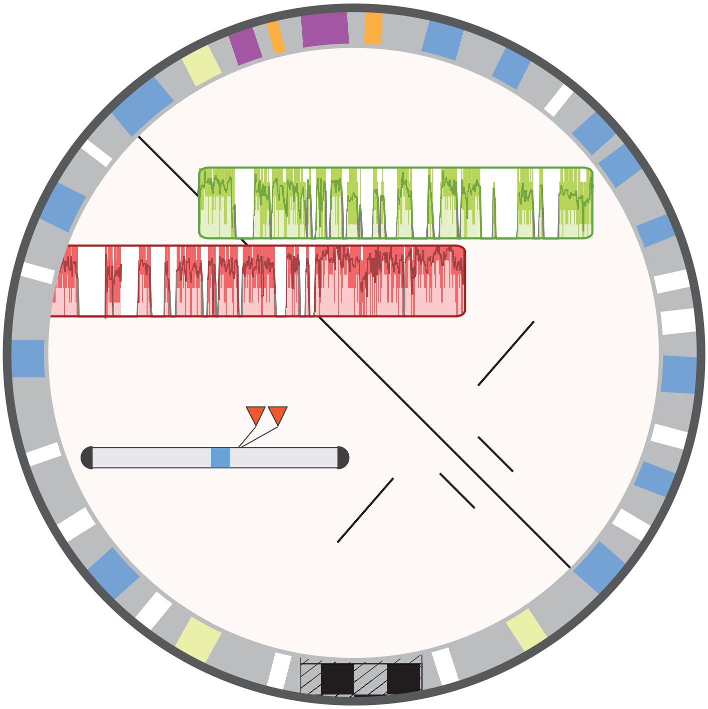
Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2021:70
Faculty of Natural Resources and Agricultural Sciences

Genomics of *Plasmodiophora brassicae*

Suzana Stjelja Arvelius



Genomics of *Plasmodiophora brassicae*

Suzana Stjelja Arvelius
Faculty of Natural Resources and Agricultural Sciences
Department of Plant Biology
Uppsala

SLU
SWEDISH UNIVERSITY
OF AGRICULTURAL
SCIENCES

DOCTORAL THESIS

Uppsala 2021

Acta Universitatis Agriculturae Sueciae
2021:70

Cover: A resting spore of *Plasmodiophora brassicae* and its genomic features. Illustration.
(photo: Suzana Stjelja Arvelius)

ISSN 1652-6880
ISBN (print version) 978-91-7760-817-2
ISBN (electronic version) 978-91-7760-818-9

Print: SLU Service/Repro, Uppsala 2021

Genomics of *Plasmodiophora brassicae*

Abstract

Plasmodiophora brassicae is a soil-borne pathogen that infects roots of plants in *Brassicaceae* and causes enlarged roots or clubs known as the clubroot disease. Details of its complex life cycle and particularly the molecular basis of its strategies to master defenses and alter metabolism of plant hosts are still largely unknown.

The general aim of this thesis was to enhance the genomic knowledge on *P. brassicae* and its intimate relationship with a plant host. We developed a protocol for extraction of large amounts of high-quality DNA from pathogen resting spores that allowed us to apply long-read PacBio sequencing. *De-novo* assembly of the *P. brassicae* e3 strain generated a 25.2 Mb nuclear genome and a mitochondrial genome (114.6 kb). Twenty nuclear contigs were assembled of which 13 from telomere-to-telomere, thanks to the resolution of highly repetitive sequences. As much as 11.5% of the genome was assigned to repetitive sequences, a higher proportion than previously estimated. The most abundant were transposable elements (TEs) such as *Copia* and *Gypsy* and unclassified interspersed repeats. They were particularly clustered in telomeric, sub-telomeric and large regions with complex structural variation found on each contig. Among 9,231 predicted protein-coding genes in the nuclear genome, we identified 314 small, secreted proteins as *P. brassicae* effector candidates. They were distributed along all contigs. TEs and unclassified repeats were found within a 3 kb distance from more than a third of the predicted effectors. We further detected enrichment of the effector candidates with a motif rich in valine, leucine and alanine amino acids. Annotation of the circular mitochondrial genome revealed a compact and complex sequence organization with intron-rich genes, a new splicing mechanism and a previously not resolved 12 kb repetitive region.

Our findings and the new genome sequences, currently representing the only *P. brassicae* long-read assembly, form a valuable resource for comparative and functional analyses.

Keywords: chromosome, clubroot, effector, genome, PacBio long-read sequencing, *Plasmodiophora brassicae*, repetitive sequences.

Author's address: Suzana Stjelja Arvelius, Swedish University of Agricultural Sciences, Department of Plant Biology, Uppsala, Sweden

Studier av *Plasmodiophora brassicae* genom

Sammanfattning

Plasmodiophora brassicae är en organism som orsakar klumprotsjuka på korsblommiga grödor och deras vilda släktingar. Denna växtpatogen tillhör Rhizaria, en till stora delar outforskad organismgrupp. Därutöver är infektionscykel och samspelet med värdväxten i stora delar okänd.

Syftet med avhandlingen var att via förbättrad kunskap om genomen i *P. brassicae* försöka klargöra vad som sker under infektionsprocessen. Initialt utvecklade vi en metod för att isolera DNA med hög kvalitet. Detta möjliggjorde att vi kunde använda sekvenseringstekniker som resulterar i långa sekvenslängder. Därigenom kunde ett förnyat kärngenom på 25,2 Mb och ett relativt stort genom representerande mitokondrien på 114,6 kb sammanfogas. Det sammanfogade nukleära genomet bestod av 13 kromosomer avgränsade med telomerer i var sin ände, samt 7 kromosombitar med noll eller en telomer-sekvens. Så mycket som 11,5% av kärngenomet bestod av repetitiva sekvenser, en högre andel än tidigare uppskattningar. Mest frekvent förekommande var transposoner som *Copia* och *Gypsy* samt oklassificerade repetitiva sekvenser. De återfanns framför allt i telomer-regionerna samt i regioner med stora strukturella avvikelser. Vi använde därefter genominformationen för att prediktera effektor-kandidater, det vill säga gener som underlättar för patogenen att angripa och utvecklas i värdväxten. Totalt 314 proteiner i den kategorin identifierades varav en delmängd anrikade med sekvenser bestående av aminosyrorna valin, leucin och alanin. Mitokondriegenomet hade en cirkulär och kompakt struktur, med intron-rika gener som använder en alternativ splitsningsmekanism. Därutöver identifierades en 12 kb lång repetitiv region som tidigare inte observerats i mitokondriegenomet.

Den nya genominformationen framtagen i detta projekt har skapat en värdefull grund för framtida analyser och ökad förståelse över hur *P. brassicae* fungerar i samspelet med värdväxten.

Nyckelord: effektor, klumprotsjuka, kromosom, kärngenom, mitokondriegenom, *Plasmodiophora brassicae,* repetitiva sekvenser.

Author’s address: Suzana Stjelja Arvelius, Swedish University of Agricultural Sciences, Department of Plant Biology, Uppsala, Sweden

Dedication

A seven-year-old girl. A teacher who believed in her.
To all my teachers for their care, commitment and knowledge.

"Science, my lad, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth."
Jules Verne

In memory of my father.

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I. Mehrabi S*, **Stjelja S***, Dixelius C (2018). Root gall formation, resting spore isolation and high molecular weight DNA extraction of *Plasmodiophora brassicae*. *Bio-Protocol, Bio-101*: e2864.

II. **Stjelja S**, Fogelqvist J, Tellgren-Roth C, Dixelius C (2019). The architecture of the *Plasmodiophora brassicae* nuclear and mitochondrial genomes. *Scientific Reports* 9, 15753.

III. **Stjelja Arvelius S**, Persson Hodén K, Fogelqvist J, Dixelius C. *Plasmodiophora brassicae*: Repetitive sequences, effectors and similarities to *Plasmodium* proteins. (manuscript)

Paper I is reproduced by following the Bio-Protocol License to Publish. Paper II is reproduced with the permission from Springer Nature. Paper III is reproduced as an authors' version.

Additional publications
Stjelja Arvelius S, Dixelius C. *Plasmodiophora brassicae*, a challenging Plasmodiophorid. (manuscript)

* These authors contributed equally

The contribution of Suzana Stjelja Arvelius and co-authors to the papers included in this thesis was as follows:

I. **SS**, SM and CD planned the project. **SS** generated *Brassica* plant materials, developed and optimized experimental procedures, performed spore isolations and DNA extractions. SM generated *Arabidopsis* plant materials, validated and standardized experimental procedures, performed spore isolations and DNA extractions and microscope work. SM, **SS** and CD wrote the manuscript.

II. **SS**, JF and CD planned the project. JF and CT-R performed bioinformatic analyses, including nuclear genome assembly and annotation. **SS** generated *Brassica* plant materials, performed spore isolations and DNA extractions and analyzed and interpreted mitochondrial genome sequencing data. **SS**, JF and CD wrote the manuscript with contribution from CT-R.

III. **SSA** and CD planned the project. **SSA** analyzed and interpreted contig variation, repetitive sequences and effector predictions. KPH made homolog searches and parasite genome comparisons. JF made input on *P. brassicae* genomic data. **SSA** and CD wrote the manuscript.

Abbreviations

aa	Amino acid
nt	Nucleotide
SC	Synaptonemal complex
SSU-rDNA	Small subunit ribosomal DNA

1. Introduction

Why would anyone study a eukaryotic microbe, not only inaccessible while hidden in soil or inside a plant but also impossible to grow in a laboratory? In case of *Plasmodiophora brassicae* there are at least two reasons: it is a plant pathogen of global agricultural importance and an organism with highly successful strategies of survival in dual environments.

As a soil-borne pathogen, *P. brassicae* infects roots of *Brassica* plants and causes characteristic enlarged clubs. Disease is therefore called clubroot and is recognized through the world due to its major impact on production of *Brassica* crops (Dixon, 2014). By reducing quality and yield of oil and vegetable crops, clubroot causes large economical losses (Dixon, 2009a). Furthermore, *P. brassicae* is almost impossible to control since its resting spores are extremely robust and can survive in soil nearly two decades (Wallenhammar, 1996). Without chemical control and with the easily disseminated pathogen that can overpower resistant cultivars (Strelkov et al., 2016) infested soils are a complex and lasting challenge for *Brassica* farmers all over the world, including Sweden (Wallenhammar et al., 2014).

To produce next generation of resting spores, *P. brassicae* requires a living plant host. Through a complex life cycle, it masters host defenses and alters host metabolism together with a plethora of other plant physiological processes (Ludwig-Müller et al., 2009; Malinowski et al., 2019). By all these changes, *P. brassicae* provides necessary nutrients and a suitable habitat for its own growth and propagation.

Many details of the *P. brassicae* lifestyle are still unknown, including cellular and molecular mechanisms. This is because plant tissues and soil are challenging environments for detail observations of the pathogen life cycle. Moreover, they obstruct isolation of larger amounts of high-quality pathogen DNA, free from host and soil microbial contamination and various inhibitors.

The first genome assembly and transcriptome of *P. brassicae* (Schwelm et al., 2015) was a breakthrough followed by a larger amount of clubroot genomic data generated in recent years. Further advancements are however slowed down because nearly half of the *P. brassicae* predicted proteins have unknown function or lack functional annotation. Comparative studies are extremely limited because whole-genome data are available for only a handful of species within Rhizaria, a large group where *P. brassicae* taxonomically belongs.

When analyzing *P. brassicae* genome data, one of the challenges is to distinguish between evolutionary broadly conserved sequences and sequence variation that is unique at the species and within-species level. Therefore, I have chosen to start this thesis introduction by providing a glimpse of the astonishing diversity of Rhizaria. Next, common characteristics of plasmodiophorids are highlighted in combination with the *P. brassicae* biology. Furthermore, current genomic resources are summarized and pathogen virulence factors introduced. Review articles are cited to provide additional information.

1.1 Taxonomic home of plasmodiophorids

1.1.1 Rhizaria

Plasmodiophora brassicae is a member of Rhizaria, a species-rich supergroup of eukaryotic microorganisms or protists. This extremely diverse group was assembled two decades ago (Cavalier-Smith et al., 2002). Rhizaria are nowadays well established as sisters of the kingdoms Alveolata and Stramenopila (Burki, 2014; Cavalier-Smith et al., 2018). The association of Stramenopila, Alveolata and Rhizaria is frequently described with the acronym SAR (Burki et al., 2007) as illustrated in **Figure 1**. SAR comprises tremendous diversity of eukaryotic species, many of which are understudied (Grattepanche et al., 2018). By inferring its sister group consisting of free-living flagellates called telonemids, SAR was updated to TSAR (Strassert et al., 2019). This and similar discoveries that shaped the current resolution of the tree of eukaryotes (**Figure 1**) are reviewed by Keeling and Burki (2019) and Burki et al. (2020).

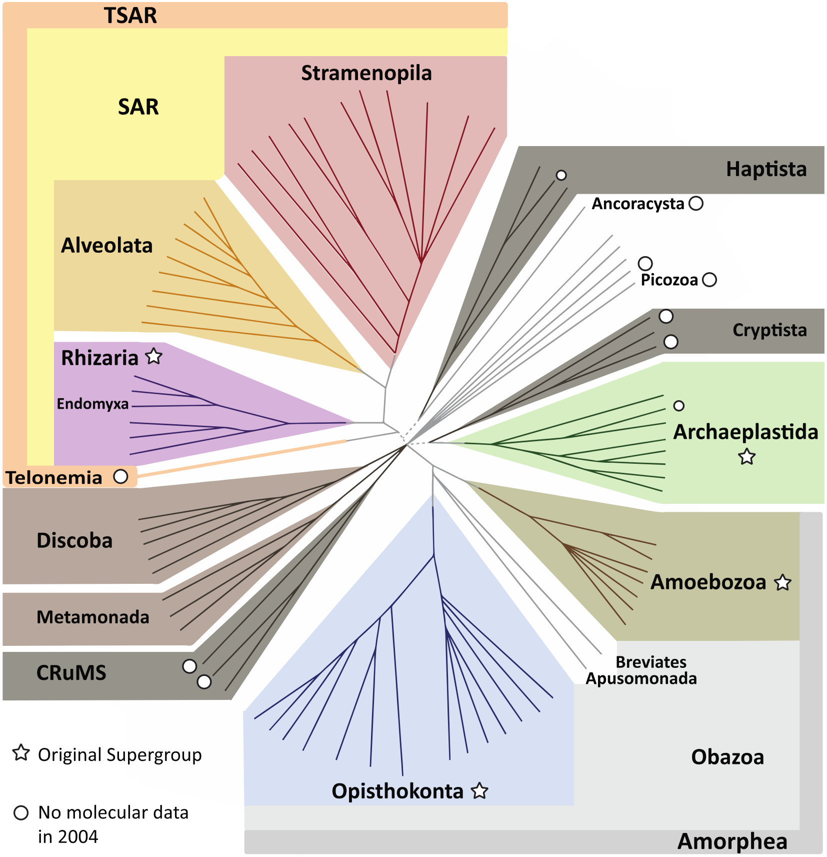


Figure 1. The eukaryotic tree of life. Adapted from Keeling and Burki (2019) and Burki et al. (2020).

Rhizaria encompass two phyla: Cercozoa and Retaria. Their internal evolutionary relationships have been re-evaluated multiple times (Cavalier-Smith et al., 2018) and new findings may bring further refinements. Members of Rhizaria were almost exclusively defined by molecular phylogenetic analyses initially based on the small subunit ribosomal DNA (SSU-rDNA) and later supported by multi-gene phylogenies (Burki et al., 2006; Sierra et al., 2016).

The oldest reliable fossil record of Rhizaria (Retaria, Foraminifera) dates to the Cambrian, 545 million years ago (mya; McIlroy et al., 2001). A molecular phylogeny with multiple fossil calibrations suggested that Foraminifera emerged between 650 and 900 mya (mean 750 mya) in the Neo-Proterozoic (Groussin et al., 2011). Thus, the origin of Foraminifera was placed about 200 mya before the fossil records. Despite a number of molecular clock estimates, the consensus for the origin of Rhizaria has not yet been reached (see Cavalier-Smith et al., 2018).

Rhizarians are morphologically one of the most heterogenous groups: 17 classes and 63 orders were catalogued by Ruggiero et al. (2015) and number of members is increasing (Elliott et al., 2019; Hittorf et al., 2020). A large variety of amoebae, flagellates, ameboflagellates, parasites and amoeboid algae are members of Rhizaria. Majority of these organisms are heterotrophic, free-living, non-photosynthetic and many are hosting microalgal symbionts. Although few ancestrally shared characteristics (synapomorphies) have been proposed (see Cavalier-Smith et al., 2018) a set of distinctive biological criteria that unify all Rhizaria and exclude other eukaryotes is difficult to define. One of the widespread features among Rhizaria are pseudopodia. These extensions of cell membrane are used for motion and for feeding by capturing and engulfing a prey. Pseudopodia can have net-like (reticulose) or thread-like (filose) appearances which closely resemble roots. Due to such resemblance and *rhíza* (the Greek word for "root") the supergroup name "Rhizaria" was minted. Extensive duplications of actin, myosin and tubulin genes and Rhizaria-exclusive protein paralogs were identified as genetic novelties that shaped evolution of the pseudopodia and cytoskeleton in these organisms (Krabberød et al., 2017).

Rhizarians are cosmopolitans found in a wide range of habitats, from soil and freshwater to open ocean waters, ice and sediment. In soil habitats, spatial and temporal distribution of Rhizaria (Cercozoa) is influenced by biotic and abiotic factors (Fiore-Donno et al., 2019). Due to intensification of agriculture, cercozoan plant pathogens are abundant in grasslands and rare or absent from forests (Fiore-Donno et al., 2020). An abundance of oceanic Rhizaria, much greater than previously estimated, is reported by a global survey (Biard et al., 2016). With extensive sampling and non-destructive *in situ* imaging methods, the survey enabled collection of vast amounts of biological data. This was particularly beneficial for plankton communities with fragile morphologies, including rhizarians which are often under-sampled or damaged by standard sampling methods. Many of planktonic Rhizaria are large (>600 µm) and form mineral skeletons consisting of calcium carbonate, silica or strontium sulfate. These skeletal structures often remain well-preserved in marine and ocean sediments. By removing carbon from surface waters to sea and ocean floors, planktonic Rhizaria are essential for the biological carbon pump (Guidi et al., 2016). They increase rates of gravitational sinking for carbon-containing particles and substantially contribute to carbon export. Some of the rhizarian taxa play important roles in the marine silica cycle by utilizing silica acid from seawater to build their skeletons (Biard et al., 2018). These siliceous rhizarians significantly contribute to global production of biogenic silica or opal and its export from surface waters to sediments (Llopis Monferrer et al., 2020).

1.1.2 Plasmodiophorids

Within the supergroup of Rhizaria and its phylum Retaria there are **Endomyxa** (**Figure 1**) which include the **Phytomyxea** class (Cavalier-Smith et al., 2018; Burki et al., 2020). Members of Phytomyxea or phytomyxids are intracellular parasites that live inside their hosts (Neuhauser et al., 2010). Because they completely depend on the host tissues for propagation, phytomyxids are so called obligate biotrophs. They can be found in any habitat worldwide, residing in various plants, oomycetes, diatoms and brown algae.

Phytomyxea further branch on **Plasmodiophorida** found in soil/fresh water and **Phagomyxida** from marine environments (Bulman et al., 2001; Neuhauser et al., 2014). These two orders currently encompass 12 genera and 42 species (Neuhauser et al., 2010; Neuhauser et al., 2011; Murúa et al., 2017). In a recent global survey, based on environmental samples and SSU-rDNA sequences, 18 new clades were generated within Plasmodiophorida (Hittorf et al., 2020). Sixteen of these clades grouped within known genera and two were novel, previously uncharacterized clades.

Before molecular studies robustly supported their classification within Rhizaria, Plasmodiophorida have experienced multiple taxonomic re-classifications. They were considered as fungi as well as slime molds and protozoa and were called Plasmodiophorales and Plasmodiophoromycetes (see Braselton, 1995; Neuhauser et al., 2010).

In contrast to fungi and oomycetes, plasmodiophorids do not exhibit nutrient-regulated (filamentous) growth and cannot be cultivated on media without their hosts. The characteristic that defines plasmodiophorids is cruciform nuclear division with a cross-like appearance of the mitotic apparatus at metaphase (**Figure 2a**) (Dylewski et al., 1978). Several other features are common among plasmodiophorids including zoospores with two anterior flagella (**Figure 2b**), intracellular forms with multiple nuclei called plasmodia (**Figure 2c**) and uninucleate resting spores (**Figure 2d**) with thick, resistant walls containing chitin (Neuhauser et al., 2010). Many of the plasmodiophorids induce increased number of cells (hyperplasia) and cell enlargement (hypertrophy) in tissues of their hosts, causing formation of galls or clubs. Moreover, plasmodiophorids are recognized for their ability to act as virus vectors. There are about 20 known, diverse species of plasmodiophorid-transmitted viruses. More details about these positive-sense single-stranded RNA viruses can be found in the reviews by Rochon et al. (2004) and Tamada and Kondo (2013).

Several pathogens of global agricultural and economic importance are members of plasmodiophorids. *P. brassicae* is the most well-known plasmodiophorid, causing clubroot on *Brassica* plants such as cabbages and oilseed rape (Dixon, 2009a). *Spongospora subterranea* is the causative agent of powdery scab on potatoes and the vector for *potato mop-top virus* (Falloon et al., 2016; Zhai et al., 2020). *Spongospora nasturtii* causes crook root on watercress and transmits *watercress yellow spot virus* (Tomlinson, 1958; Tomlinson and Hunt, 1987). The species was recently renamed to *Hillenburgia nasturtii* (Hittorf et al., 2020). Plasmodiophorids which do not cause galling or other direct disease symptoms are found among *Ligniera* (Neuhauser and Kirchmair, 2009) and *Polymyxa* species which transmit a number of viruses. *P. betae* is the vector for four viruses, including *beet necrotic yellow vein virus* that causes rhizomania, one of the most important diseases of sugar beets (McGrann et al., 2009; Tamada and Asher, 2016). *P. graminis* transmits several viruses causing diseases on cereal crops such as wheat, oat and rice (Kanyuka et al., 2003). These and other less known plasmodiophorids are reviewed by Bulman and Neuhauser (2016). Many details of plasmodiophorid complex life cycles are unknown and most information is based on observations from *P. brassicae* (**chapter 1.2.2**). Valuable information can be found, together with an impressive gallery of microscope images on the Plasmodiophorid Home Page.

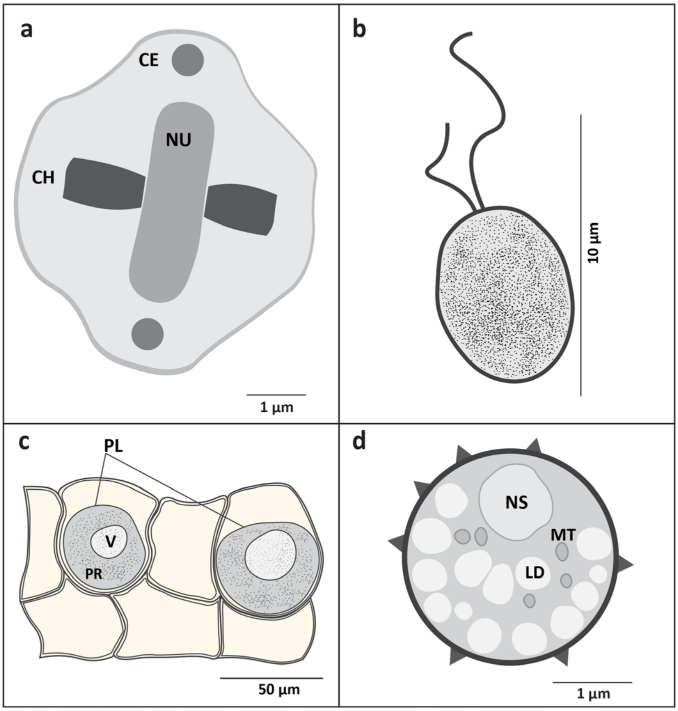


Figure 2. Common characteristics of plasmodiophorids. a. Cruciform nuclear division (CE-centrioles, CH-chromatin, NU-nucleolus, adapted from Bulman and Braselton, 2014), **b**. **Biflagellate zoospore** (adapted from Tomlinson, 1958), **c**. **Plasmodia** (transversal root sections with PL-plasmodium, PR-protoplasm, V-vacuole, adapted from Tomlinson, 1958), **d. Resting spore** (LD-liquid droplet, MT-mitochondria, NS-nucleus, adapted from Bulman and Braselton, 2014 and Bi et al., 2016).

Phagomyxids, a sister clade to plasmodiophorids, are obligate biotrophs in marine environments (Neuhauser et al., 2014). *Maullinia* are recognized as parasites of brown algae, including *M. ectocarpii* (Parodi et al., 2010) and *M. braseltonii* which causes galls on infected bull kelp (Murúa et al., 2017). *Phagomyxa bellerochae* and *P. odeontellae* parasitize on marine diatoms (Schnepf et al., 2000). A novel phagomyxid parasite has been described to incite galls on root hairs of eelgrass and to share a high sequence similarity with *Plasmodiophora diplantherae* (Elliott et al., 2019). *P. diplantherae*, recently re-named to its former name *Ostenfeldiella diplantherae* (Hittorf et al., 2020) is known to cause shoot galls in seagrasses (Walker and Campbell, 2009). *Marinomyxa halophilae* and *M. marina* represent gall-forming parasites of seagrasses (Kolátková et al., 2020; Kolátková et al., 2021).

1.1.3 Rhizarian genomic information

Rhizaria are one of the least understood groups of eukaryotes and there are large gaps between their remarkable diversity and available genomic data (del Campo et al., 2014; Sibbald and Archibald, 2017; Grattepanche et al., 2018). Many of these organisms remain un- or under-sampled due to their fragile, easily damaged morphologies (marine species) or because they are shielded by resilient, thick-walls of resting spores (soil-borne species). Furthermore, they are difficult or impossible to cultivate on media. These characteristics in combination with complex habitats (water, soil and host intracellular space) pose significant challenges for isolation of their pure DNA and RNA.

Most of sequence information for Rhizaria derives from the SSU-rDNA and other genes, widely used for single- and multi-gene phylogenies. A large number of these sequences is revealed by sequencing of the SSU-rDNA from environmental samples (Biard et al., 2016; Fiore-Donno et al., 2019; Hittorf et al., 2020). There are however rhizarians such as foraminiferans which cannot be detected by standard environmental surveys due to insertions and substitutions in their SSU-rDNA sequences (Pawlowski, 2000). Taken together, the actual diversity and abundance of Rhizaria is most likely much greater than estimated and many sequences are yet to be discovered and classified.

Whole genome sequencing projects commonly require larger amounts of high-quality DNA that is free from host and microbial contamination and inhibitors. Consequently, genome data for Rhizaria are rare and still biased towards only few photosynthetic members and parasites (Burki and Keeling, 2014). Only 15 species with genome sequences (including nuclear and/or mitochondrial, plastid, nucleomorph and chromatophore genomes) are currently listed for Rhizaria by the Taxonomy database (Schoch et al., 2020) at the National Center for Biotechnology Information (NCBI) accessed via https://www.ncbi.nlm.nih.gov on 07.09.2021. These records belong to five photosynthetic chlorarachniophyte algae, two photosynthetic euglyphid amoebae, three foraminiferans, three plasmodiophorid plant endoparasites (**Table 1**), a cercomonad flagellate and to an unclassified member. Overall, the NCBI Assembly database (Kitts et al., 2016) lists 62 genome assemblies (https://www.ncbi.nlm.nih.gov/assembly/?term=Rhizaria). More than one nuclear and/or organellar genome can be included in an assembly record.

Transcriptomes of Rhizaria are valuable complements to the genome records. These data are generated by RNA-sequencing with the goal to identify expressed genes. Examples include transcriptomes generated by the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP, Keeling et al., 2014), transcriptomes of soil-borne plasmodiophorid organisms (**Table 1**) and single-cell transcriptomes (Krabberød et al., 2017).

Among 42 known phytomyxid species (Neuhauser et al., 2010; Neuhauser et al., 2011; Murúa et al., 2017) the only organisms with available whole genome data are the three plasmodiophorids (**Table 1**). These phytomyxids were introduced into the genomic era by sequencing a nuclear genome of the clubroot pathogen, *P. brassicae* (Schwelm et al., 2015). Furthermore, transcriptomes of *P. brassicae* and *S. subterranea*, the causing agent of potato scab, were reported in the same study. Since then, a larger amount of plasmodiophorid data has been generated, including genome assemblies of *S. subterranea* (Gutiérrez et al., 2016; Ciaghi et al., 2018) and *P. betae*, the vector of rhizomania in sugar beets (Decroës et al., 2019). *P. brassicae* remains the most studied plasmodiophorid with 50 genome assemblies from multiple strains listed at the NCBI Genome database (https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/38756/). For additional details on these data, see **chapter 1.2.5.**

There is a pronounced need for whole genome assemblies among phytomyxids and Rhizaria as well as protists in general. More of high-quality genome data will enable direct comparisons and deeper analysis of genomes as well as facilitate use of other sequence resources such as organism and single-cell transcriptomes, metagenomes and metatranscriptomes. Development of novel culture methods and application of culture independent techniques (single-cell genomics) in combination with joint international efforts (Cheng at al., 2018; Miao et al., 2020) and deposition of data in public databases will greatly benefit genomics of Rhizaria.

Table 1. Plasmodiophorid genomic data.

Species	Genomic data	Number of assemblies	Sequencing technology[a]	Genome size	Reference
Plasmodiophora brassicae	Nuclear genome	50[b]	454 (2)	24.04 - 25.25 Mb	Rolfe et al., 2016
			Illumina/454 (2)		Schwelm et al., 2015; Rolfe et al., 2016
			Illumina (45)		Bi et al., 2016; Daval et al., 2019; Sedaghatkish et al., 2019
			PacBio (1)		Stjelja et al., 2019
	Mitochondrial genome	4[c]	Illumina (3)	93.64 - 114.66 kb	Bi et al., 2016; Rolfe et al., 2016; Daval et al., 2019
			PacBio (1)		Stjelja et al., 2019
	Transcriptome	3	Illumina (3)		Schwelm et al., 2015; Bi et al., 2016; Daval et al., 2019
Spongospora subterranea	Nuclear genome	1	Illumina	28.08 Mb	Ciaghi et al., 2018
	Mitochondrial genome	1	454	37.69 kb	Gutiérrez et al., 2016
	Transcriptome	2	Illumina (2)		Schwelm et al., 2015; Baloth et al., 2021
Polymyxa betae	Nuclear genome	1	Illumina	27.08 Mb	Decroës et al., 2019

a = number of assemblies specified in the brackets.

b = NCBI Genome database, accessed 07.09.2021.

c = two mitochondrial genome assemblies and two mitochondrial sequences within whole genome projects.

1.2 *Plasmodiophora brassicae* Воро́нин

1.2.1 Clubroot

P. brassicae infects plants in the mustard family (*Brassicaceae*, *Cruciferae*) and causes characteristic enlargement of roots or root clubs (**Figure 3a, b**). The disease is therefore called clubroot (for common names in a number of languages see Dixon, 2009a). Clubroot is the most widely known disease caused by a plasmodiophorid pathogen. It has a major impact on global production of *Brassica* such as oil crops (e.g. oilseed rape), vegetable crops (e.g. cabbages), spices (e.g. mustard seeds) and forage used as animal feed (Dixon, 2014). By reducing quality and yield in range from 10 to 100%, clubroot causes large losses in *Brassica* crop production, lowers the value of infected land and involves significant costs for management strategies (Dixon, 2009a; Howard et al., 2010).

Clubroot is today established throughout the world, particularly in *Brassica* growing countries in Europe, Asia (India, Nepal, China, Japan, Korea), Australia, New Zealand, Canada and Latin America (Dixon, 2009a; Botero et al., 2019). Intensification of agriculture and popularization of *Brassica* crops are important factors behind the clubroot widespread presence and a steady increase of its incidence. There are records indicating that clubroot has been recognized in Europe from the 13th century and some estimates suggest even from the 4th century and Roman times (see Dixon, 2009a; Dixon, 2014). The causative agent of the disease was not identified until the late 19th century. In areas around Saint Petersburg, a large outbreak of clubroot destroyed cabbage crops which were essential as winter food and feed supplies. After the outbreak, a Russian biologist Михаи́л Степа́нович Воро́нин started a quest for the cause of the disease and suitable control measurements. In 1878 Воро́нин described *Plasmodiophora brassicae*, a previously unknown organism. By identifying parts of the pathogen complex life cycle, Воро́нин established the casual correlation between *P. brassicae* and symptoms of cabbage hernia or clubroot (Neuhauser et al., 2010).

P. brassicae is a soil-borne pathogen and its resting spores are dispersed via infested soil and contaminated equipment and tools, vehicles, clothes and footwear, animals and livestock manure (Chai et al., 2016), root nematodes and insects as well as by water (Datnoff et al., 1987; Howard et al., 2010) and windblown dust (Rennie et al., 2015).

Resting spores of *P. brassi*cae are extremely resilient and can survive in soil nearly two decades (Wallenhammar, 1996). Furthermore, they are easily disseminated between fields and have a high reproduction rate, with an estimated 100,000-fold inoculum increase per generation (Diederichsen et al., 2009). Since no commercial chemical agents are available and *P. brassicae* can overpower resistant cultivars (Strelkov et al., 2016; Cao et al., 2019) management of infested soils is complex. Different control measures are employed such as rotation of susceptible *Brassica* crops and non-host crops, use of decoy hosts, application of agricultural lime and wood ash and use of calcium cyanamide fertilizer alone and in combination with bio-fertilizers (Dixon, 2009b; Howard et al., 2010; Dixon, 2017).

Characteristic symptoms of clubroot are enlarged roots called clubs or galls (**Figure 3a**). The clubs serve as a nutrient sink and a suitable habitat, necessary for *P. brassicae* growth and propagation. Due to redirection of nutrients (particularly carbohydrates) from the plant upper parts towards the clubroots, infected plants often show reduced growth. Other above-ground symptoms such as wilting and yellowing are common since infected plants have lowered capacity for water and nutrient uptake (**Figure 3a**). Soft and mushy clubroots disintegrate at the end of disease cycle, releasing a huge number of *P. brassicae* resting spores into soil. Clubroot symptoms and time of their incidence depend on host species and conditions such as temperature, soil humidity and spore load. For example, clubroots of *Brassica rapa* (Chinese cabbage) plants grown at 22°C and 60% relative humidity start to decompose 8 to 10 weeks after infection with *P. brassicae*. For gentle plants of *Arabidopsis thaliana* (thale cress), grown in the same conditions, this period is 4 to 5 weeks long.

Through various strategies *P. brassicae* masters host defenses and alters host metabolism together with a plethora of other plant physiological processes. Details about changes in plant metabolism caused by the clubroot pathogen are described by Ludwig-Müller et al. (2009) and information about gall architecture and other cellular changes in infected plants can be found in the review by Malinowski et al. (2019).

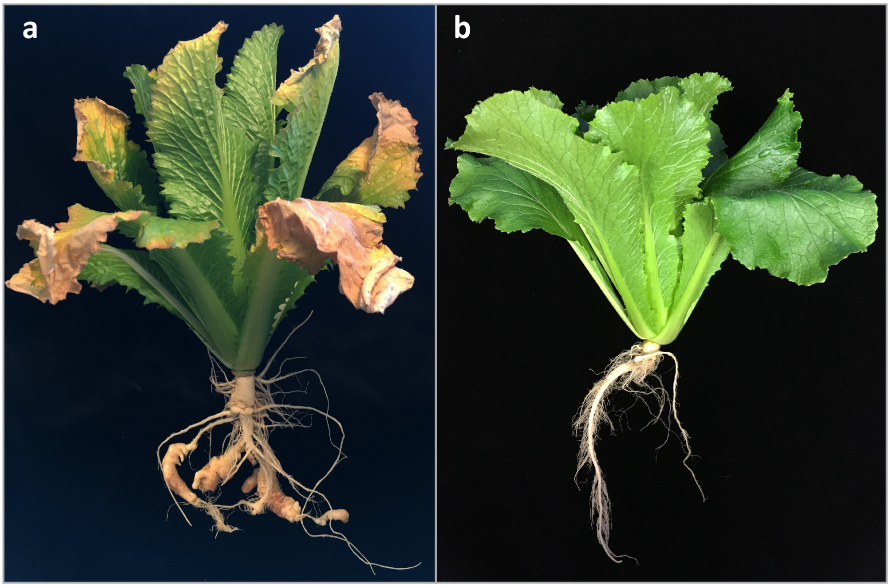


Figure 3. Chinese cabbage (*Brassica rapa*) **a.** Plant infected with *Plasmodiophora brassicae,* displaying above- and below-ground clubroot symptoms. **b**. Healthy plant. Photo: Suzana Stjelja Arvelius.

1.2.2 Life cycle of the clubroot pathogen

Due to its obligate biotrophic nature, *P. brassicae* has a life cycle shared between soil and a plant host. As such, life events are complicated for observations and many details are still unknown. Broadly, the life cycle can be divided in two periods: a dormant period with resting spores in the soil (**Figure 4.1**) and an active, mostly intracellular period (**Figure 4.2 to 4.9**) which includes primary and secondary infection of host tissues.

Transition towards the active state is initiated by germination of a **resting spore** (**Figure 4.1**) (MacFarlane, 1970; Tanaka et al., 2001). Exact nature of factors that trigger germination is unknown, signals from host and non-host root exudates as well as spore-intrinsic mechanisms have been suggested (see Dixon, 2009b). Each resting spore develops into a uninucleate **primary zoospore** with two flagella (**Figure 4.2**). The primary zoospore is motile and “swims” in the soil. In contrast to its otherwise well-protected life (by a spore wall or host tissues), *P. brassicae* is in this phase fragile and exposed to various biotic and abiotic factors. It is uncertain to which extend environmental factors (soil temperature, pH, chemical composition and texture) affect motility or whether motility is a passive or an active, flagella-driven process (Dixon, 2014). Higher soil moisture with presence of free water between soil particles (Gravot et al., 2016) and temperatures in range 20-25°C (Sharma et al., 2011) are recognized as favorable for the zoospore motility, germination and clubroot development. When in a vicinity of a compatible host, the primary zoospore starts host infection by attaching to the surface of a root hair. Next, the flagella are retracted and the **zoospore encysts** (**Figure 4.3**) and develops special internal structures called Rohr and Stachel (Aist and Williams, 1971; Williams et al., 1973; see Schwelm et al., 2018). By using these structures, the **zoospore penetrates the root hair wall** (**Figure 4.4**) and injects its own cellular content with a single nucleus, into the host cell. This leads to development of a primary plasmodium inside the root hair cell. After several mitotic, cruciform nuclear divisions a **multinucleate plasmodium** is formed (**Figure 4.5**) and then cleaved into **zoosporangia** (**Figure 4.6**). From each zoosporangia several **secondary zoospores** (**Figure 4.7**) are hatched and released from the root hair into the soil (Bulman and Braselton, 2014).

A secondary, biflagellate zoospore initiates secondary infection by penetrating root cortical cells where it develops into a secondary plasmodium. Next, after several nuclear divisions a **multinucleate plasmodium** (**Figure 4.8**) is formed. This phase is associated with extensive proliferation and enlargement of the host cells and development of the root clubs. At the end of the life cycle, the multinucleate plasmodium cleaves and matures into uninucleate **resting spores** (**Figure 4.9**) which are released into the soil (**Figure 4.1**).

There are still many unknowns about the *P. brassicae* life cycle. Can zoospores move (actively and/or passively) through plant tissues? Can primary zoospores directly initiate secondary infection? Furthermore, do potential meiosis and genetic recombination occur? Tommerup and Ingram (1971) suggested that a major part of the *P. brassicae* life cycle, including resting spores, is haploid while a short diploid phase occurs after nuclear fusion of secondary zoospores. Recently, a conjugation of two secondary zoospores and formation of a diploid zygote was described, indicating the sexual phase (Liu et al., 2020). However, absence of sexual recombination was also reported (Fähling et al., 2004) and no mating type genes have been found in the *P. brassicae* first genome draft (Schwelm et al., 2015).

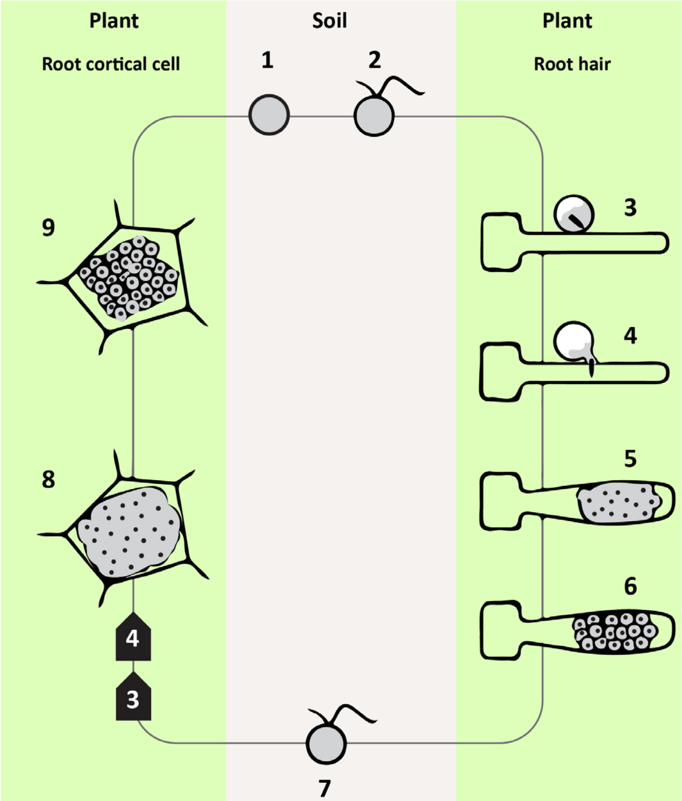


Figure 4. Life cycle of *Plasmodiophora brassicae.* Adapted from Bulman and Neuhauser (2016) and Williams et al. (1973). Life-stages (1 to 9) are described in the main text.

Details about the complex life of *P. brassicae* can be found in the review by Kageyama and Asano (2009). Schuller and Ludwig-Müller (2016) provide an extensive summary of numerous microscopic techniques and staining methods used for visualization of *P. brassicae* in host tissues. A new, refined information about infection process, life cycle events and the sexual stage is presented by Liu et al. (2020).

1.2.3 Variation in virulence and pathotype characterization

Different pathotypes and strains of *P. brassicae* have been recognized. Distinction between these two terms is not very clear in the *P. brassicae* literature and they are often used interchangeably. In general, a strain represents a genetically distinct variant of a microorganism while a pathotype is defined according to its ability to cause disease or pathogenicity. The degree of pathogenicity is described by virulence (Shapiro-Ilan et al., 2005).

Pathotypes of *P. brassicae* differ by their ability to infect various *Brassica* hosts and by severity of the clubroot symptoms they cause. Moreover, some pathotypes have the capacity to overcome host resistance (Hatakeyama et al., 2004; Strelkov et al., 2016). A mixture of pathotypes is often found in an infested field or within a plant and even within a single clubroot (Fu et al., 2020).

Differences in virulence are utilized for pathotype characterization, a process that provides important information for *Brassica* breeders and the clubroot research community. Defining a pathotype is, however, a complex task in case of *P. brassicae* because there are several classifications systems (Karling, 1942; Ayers, 1957; Williams, 1966; Buczacki et al., 1975; Somé et al., 1996; Kuginuki et al., 1999; Hatakeyama et al., 2004; Kim et al., 2016; Strelkov et al., 2018; Pang et al., 2020). These systems differ by selected virulence criteria and by *Brassica* species and cultivars used for disease assessment. Since the assessment is based on plant phenotyping, there is also variation within a system (e.g. human and/or plant factor). Taken together, comparison of pathotypes that were characterized in different systems is complicated.

Accurate, robust and reproducible molecular diagnostic tests that would replace the labor- and time-consuming plant phenotyping are needed. Such tests would enable more efficient monitoring, quantification and management of the clubroot pathogen. Development of a DNA marker test requires identification of nucleotide sequence(s), unique to a particular pathotype. This complex task is now facilitated by comparison of whole genome sequences from several pathotypes (Jeong et al., 2018) and identification of genomic insertions (Holtz et al., 2021a). However, one of the challenges is to achieve high differentiating capacity of the tests.

Schwelm and Ludwig-Müller (2021) describe this and other challenges of the work on pathotype-specific markers and provide a comprehensive overview of the *P. brassicae* classifications systems and pathotypes. The review by Tso et al. (2021) offers a summary of technical details for several pathotyping platforms.

1.2.4 Genetic diversity of pathotypes

P. brassicae pathotypes show variation in virulence patterns but is not yet fully understood how pathotype diversity is maintained and modified in populations. Next, how similar or distinct are pathotypes based on their genetic basis? What molecular mechanisms are involved in genetic diversity of the clubroot pathogen?

In experiments with repeated exposure to several host genotypes, from susceptible to resistant, *P. brassicae* quickly adapted by causing more severe symptoms in the resistant hosts (LeBoldus et al., 2012). Similarly, more virulent *P. brassicae* pathotypes emerged after continuous planting of clubroot-resistant cultivars under greenhouse conditions (Cao et al., 2019). These and other examples of erosion and loss of clubroot resistance (Hatakeyama et al., 2004; Strelkov et al., 2016) demonstrate that *P. brassicae* can modify its virulence and pathotype diversity in response to shifts in host resistance.

Populations of the clubroot pathogen are recognized as heterogenous, complex mixtures of pathotypes (Manzanares-Dauleux et al., 2001; Strelkov et al., 2018). A population is not clearly defined in studies of *P. brassicae* and it can include spores from a single or several clubroots as well as materials from a field, a region or a country. Consequently, the size of populations is variable.

It was suggested that more virulent pathotypes could exist at initially low frequencies within a population (Sedaghatkish et al., 2019). Due to selection imposed by use of resistant hosts, they would increase in proportion and consequently emerge as dominant. A decade long, marker-based screening of more than 200 clubroots supported this assumption by revealing that highly virulent pathotypes were present before introduction of the clubroot-resistant cultivars (Holtz et al., 2021b). Such preservation of rare pathotypes within population mixtures in clubroots could be a result of balancing selection, identified to act on *P. brassicae* populations (Sedaghatkish et al., 2019). More diverse populations, maintained by balancing selection, have potential advantages when adapting to changing environmental conditions.

Estimates of genetic variation in *P. brassicae* populations vary, from high to low levels. Canadian populations were found highly diverse based on nucleotide diversity of whole-genome sequences from 43 strains, majority of which from various regions in Canada and few from China and the USA (Sedaghatkish et al., 2019). Populations from clubs in several regions in Germany (Strehlow et al., 2014) and different fields in France (Manzanares-Dauleux et al., 2001) displayed extensive differences in patterns of polymorphic DNA markers. In these three studies, observed genetic variation could not be classified according to hosts or virulence phenotypes. Moreover, detected genotypes exceeded pathotype classes in the two European populations. Clustering based on geographical distribution allowed differentiation of the German populations in two genetic classes (Strehlow et al., 2014). Opposite to the extensive variation, low genetic diversity was reported within two Canadian populations (Holtz et al., 2018). One avirulent and another virulent on resistant hosts, each of these populations was very similar and considered even clonal based on restriction site-associated DNA sequencing (RADseq). Furthermore, four of five Canadian pathotypes displayed only small differences based on whole-genome single nucleotide polymorphisms (SNPs) (Rolfe et al., 2016). All SNP profiles clustered according to the host range and geographical distribution of the pathotypes.

The genetic structure of *P. brassicae* populations seems to be shaped by balancing selection and selective pressures imposed by the host specificity and resistance range as well as by the geographical origin. Besides selective processes, the genetic variation is generally influenced by population size, genetic drift, mutation rate, migration and sexual recombination.

Populations of the clubroot pathogen are expected to have a large effective size thanks to robust, long-lived resting spores (Wallenhammar, 1996) and a high reproduction rate per generation (Diederichsen et al., 2009). Furthermore, the size of local populations can be influenced by their age and host cropping systems (Strehlow et al., 2014). Genetic drift (random loss or fixation of genetic variants) most likely is not a significant driver of diversity in *P. brassicae* since its effect is larger in smaller populations (Sedaghatkish et al., 2019). There are no estimates of mutation rate in the clubroot pathogen.

Migration between populations does occur (Sedaghatkish et al., 2019) but there is uncertainty regarding the level and effect of gene flow. As a soil-borne pathogen, *P. brassicae* reaches longer distances primarily via infested soil and contaminated equipment and water. Thus, isolated fields could have only a limited spore dispersal and low gene flow. Such scenario was suggested as an explanation for the high genetic variation found in populations from several regions in Germany (Strehlow et al., 2014). Whether sexual reproduction occurs in *P. brassicae* is not yet fully understood (**chapter 1.2.2**). Polymorphism in chromosomal size but no sexual recombination could be detected by repetitive fragments used as hybridization probes (Fähling et al., 2004). More recent estimates of recombination, based on linkage disequilibrium (LD) yielded different results in the populations that were considered as sexual based on their extensive genetic variation. Reduction of LD indicated a high recombination frequency in populations across Canada (Sedaghatkish et al., 2019). On the other hand, high LD was found in populations from German regions (Strehlow et al., 2014).

1.2.5 Genomic information

Distinction between field isolates and single spore isolates is important for genomic studies of the clubroot pathogen. This is because field samples including soil, a plant or a single clubroot can contain genetically diverse strains (Fu et al., 2020; Holtz et al., 2021b). Single spore isolates are generated by infecting a host with a single spore to multiply genetically stable progeny (Fähling et al., 2004; Heo et al., 2009). If pathotype and strain classifications are not known, field and single spore isolates are pathotyped and characterized based on their genetic diversity.

P. brassicae resting spores, often used for extraction of genomic DNA, have haploid nuclei (Tommerup and Ingram, 1971). The exact number, size and shape of chromosomes (karyotype) is still unknown. Estimates range from 5 haploid chromosomes (Tommerup and Ingram 1971; Ingram and Tommerup, 1972) and 6 to 16 chromosomal bands (Bryan et al., 1996; Ito et al., 1994; Graf et al., 2001; Graf et al., 2004) to 20 haploid chromosomes (Braselton, 1982). Polymorphism in size of chromosomes has been reported for a single spore isolate (Fähling et al., 2004) and among strains (Graf et al., 2004).

Multiple strains of *P. brassicae* have been sequenced and data from 50 genome assemblies are currently available (**chapter 1.1.3 and Table 1**). There is, however, a discrepancy in comparison with the number of genome sequences (52) described in the literature (Schwelm et al., 2015; Bi et al., 2016; Rolfe et al., 2016; Daval et al., 2019; Sedaghatkish et al., 2019; Stjelja et al., 2019). Only two (AAFC-SK-Pb3 and AAFC-SK-Pb6) of five Canadian strains (Rolfe et al., 2016) are available in the NCBI Genome database, accessed on 07.09.2021. Furthermore, two entries with similar records can be found for the Pb3 strain (AAFC-SK-Pb3 and Pb3). Taken together, there are 49 distinct genome assemblies of *P. brassicae*. These assemblies are thoroughly described in a recent review (Schwelm and Ludwig-Müller, 2021). The review provides an essential link between genome data and information about isolates and pathotypes from multiple classification systems and hosts and country of origin.

P. brassicae genome assemblies have sizes between 24.04 and 25.25 Mb (**Table 1**). The number of predicted protein-encoding genes in annotated assemblies is from 9,231 to 12,811 genes (Stjelja et al., 2019; Schwelm et al., 2015; Rolfe et al., 2016; Daval et al., 2019). Roughly a half of these genes lack functional annotation or have predicted unknown or general function.

Mitochondrial sequences are available for four *P. brassicae* strains, encompassing from 93.6 to 114.6 kb (**Table 1**) and encoding 60 to 79 predicted genes (Daval et al., 2019; Stjelja et al., 2019).

Only few transcriptomes, generated by sequencing of RNA from spores in various life-stages of *P. brassicae* are currently available (**Table 1**). Expression data from various hosts infected by *P. brassicae* are much more abundant and only few studies are cited here (Zhao et al., 2017a; Irani et al., 2018; Ciaghi et al., 2019; Olszak et al., 2019; Daval et al., 2020; Pérez-López et al., 2020). Sequencing data from several clubroot microbiome studies are available (Zhao et al., 2017b; Lebreton et al., 2019; Tian et al., 2019).

1.3 Effectors

Often described as an evolutionary arms race, plant-pathogen interactions involve a wide range of sophisticated countermeasures. Plant defense is based on innate immunity and a surveillance system with many immune receptors (Jones and Dangl, 2006). In the first line of defense, pathogen associated molecular patterns (PAMPs) are recognized by pattern recognition receptors (PRRs). PAMPs encompass conserved molecules such as proteins, oligosaccharides and lipopolysaccharide commonly found in pathogen cell walls (e.g. flagellin, peptidoglycan, cellulose, chitin). Their recognition leads to PAMP-triggered immunity (PTI). Successful pathogens can overcome PTI thanks to secretion of so-called effector proteins. Recognition of these proteins by receptors from the second line of defense (resistance or R proteins) elicits effector triggered immunity (ETI). Effectors are primarily known as virulence factors while detected effectors represent avirulence (Avr) proteins (Jones and Dangl, 2006). ETI commonly results in the hypersensitive response, a type of programmed cell death that rapidly occurs to prevent a pathogen from further invasion (Dodds and Rathjen, 2010). Modifications of these interactions and new models of plant defense systems are proposed (Wang et al., 2019; Thordal-Christensen, 2020).

Pathogen use effectors to promote fitness and virulence in various ways: by facilitating host colonization and infection, manipulating host defenses and avoiding recognition and interfering, inhibiting or mimicking host cellular reactions (Białas et al., 2018; Jaswal et al., 2020). Effectors that are secreted into the plant extracellular space are called apoplastic effectors. Cytoplasmic effectors are delivered into the plant cell where they target single or multiple subcellular compartments while some effectors have membrane-localization. Mechanisms of delivery are diverse and unknown for many effectors. Many of fungal and oomycete effectors possess short host-targeting motifs (e.g. RxLR where x represents any amino acid, DEER and LxFLAK) necessary for translocation into a plant cell. A comprehensive overview of delivery mechanisms and key cellular processes targeted by effectors from many pathogens is provided by Toruño et al. (2016).

Cysteine residues and their intramolecular disulphide bonds are recognized to play important roles for protein folding and function and effector stability in the plant apoplast (Kamoun, 2006; Stergiopoulos and de Wit, 2009). It is speculated that variable regions of the cysteine-rich proteins might allow significant changes in amino-acid sequences without affecting protein overall structure, contributing to emergence of new virulence (Povolotskaya and Kondrashov, 2010). Furthermore, conserved cysteine-rich motifs such as CHXC and CXHC might be involved in delivery of effectors into host cells (Kemen et al., 2011). Lack of cysteine richness in some effectors (Gout et al., 2006) indicate that the cysteine content criteria should be carefully applied in effector prediction pipelines (Sperschneider et al., 2015).

Because of strong selection pressure imposed by the host recognition, effectors evolve rapidly and are extremely diverse in their sequence structure (Upson et al., 2018). Some pathogen genomes can be divided into a conserved, gene-dense and a fast-evolving region. Terms “two-speed” (Raffaele and Kamoun, 2012; Dong et al., 2015) or two-compartment genomes (Frantzeskakis et al., 2019) are used to describe such architecture. Highly dynamic regions are often rich in repeat and transposable elements (TEs) and harbour effectors. In other words, effector genes seem not to be randomly distributed across the genomes. Instead, they tend to reside within the regions with higher mutation rates and sequence rearrangements. It is suggested that such organization can, by various mechanisms, accelerate effector diversification and pathogen evolution (Sánchez-Vallet et al., 2018).

A methyltransferase that methylates salicylic acid (a plant hormone involved in defense) and reduces its levels in infected roots was the first confirmed *P. brassicae* effector (Ludwig-Müller et al., 2015). After this discovery, a larger number of effector proteins was predicted and candidates for functional analysis were selected thanks to availability of genome sequences and transcriptome data (Schwelm et al., 2015; Rolfe et al., 2016; Holtz et al., 2018; Daval et al., 2020; Pérez-López et al., 2020). Among these predicted effectors, proteins with ankyrin repeats and chitin-binding domains were commonly found. However, a majority of the effector candidates lack functional annotation. Since there is no well-established transformation protocol for the clubroot pathogen, analysis of a gene function by gene knock-out and gene silencing technologies is difficult or impossible to apply. More information about functional roles of *P. brassicae* predicted effectors can be obtained by heterologous expression (expression of a gene in a host organism) as reported by several studies (Singh et al., 2018; Yu et al., 2019; Chen et al., 2019; Pérez-López et al., 2021). Additional details about *P. brassicae* effector candidates can be found in the review by Schwelm and Ludwig-Müller (2021).

2. Aims of the study

Genome studies of pathogens are greatly enhancing our understanding of their life cycles, molecular mechanisms of infection and host manipulation, genome organization and sequence features of importance for virulence. Such knowledge has a particular value for a largely understudied group of soil-borne plant pathogens. One of these enigmatic pathogens, *P. brassicae* has been in the focus of this thesis. The general aim of my work was to enhance the genomic knowledge on *P. brassicae* and its interaction with a plant host.

Specific objectives of my PhD projects included in this thesis were to:

- Develop a reproducible protocol for extraction of large amounts of high-quality DNA from *P. brassicae* resting spores,

- Apply long-read PacBio RS II sequencing technology to refine information from the nuclear and mitochondrial genomes of the *P. brassicae* e3 strain,

- Perform mitochondrial genome *de-novo* annotation and comparative and phylogenetic analyses,

- Explore nuclear intra-genomic variation and perform a deeper analysis of annotated repetitive sequences,

- Gain knowledge about *P. brassicae* effector candidates, including their genome distribution and motif enrichment.

3. Results and Discussion

The results obtained in this thesis are comprehensively described and discussed in the three papers. In this section, I will present a summary of the main results and highlight some of the findings to interlink them into a wider context. In the text, the papers are referred to by Roman numerals and the results are indicated by their position within a corresponding paper.

3.1 *P. brassicae* high-quality DNA (Paper I, II, III)

Refining genome information of *P. brassicae* e3 strain by applying long-read PacBio RS II sequencing technology was essential for this thesis. PacBio data allowed us to resolve complex, repetitive regions and obtain a telomere-to-telomere assembly of the nuclear genome and a mitochondrial sequence (Paper II and III). None of these results would have been possible without developing a protocol for extraction of large amounts of high-quality DNA from resting spores of *P. brassicae* (Paper I). Obtaining contamination free DNA from an obligate biotroph is a challenging task and it would not have been realized without previous efforts and extensive work on the clubroot pathogen (Schwelm et al., 2015). We hope that Paper I will facilitate extractions of DNA from various *P. brassicae* strains and pathotypes, with quality and quantity suitable for genome sequencing and other molecular analysis.

3.1.1 *Brassica* and *Arabidopsis* clubroots

We have grown *Brassica rapa* (Chinese cabbage) plants in soils infested with *P. brassicae* e3 strain (Fuchs and Sacristan, 1996; Klewer et al., 2001) to generate clubroots for spore isolation (Paper I, Fig. 1).

B. rapa is a robust plant and can develop relatively large clubroots. *Arabidopsis* plants are, on the other hand, quite delicate and they commonly develop tiny galls after infection with *P. brassicae* (Paper I, Fig. 2). Thus, many *Arabidopsis* galls are necessary to harvest for spore isolation and extraction of nucleic acids. Thanks to a master thesis work by Elen Lefeuvre and her efforts to fine-tune several settings from Paper I (soil spore load, seedling transplanting technique, growing conditions and harvesting time) we were able to produce larger *Arabidopsis* clubroots (**Figure 5a**).

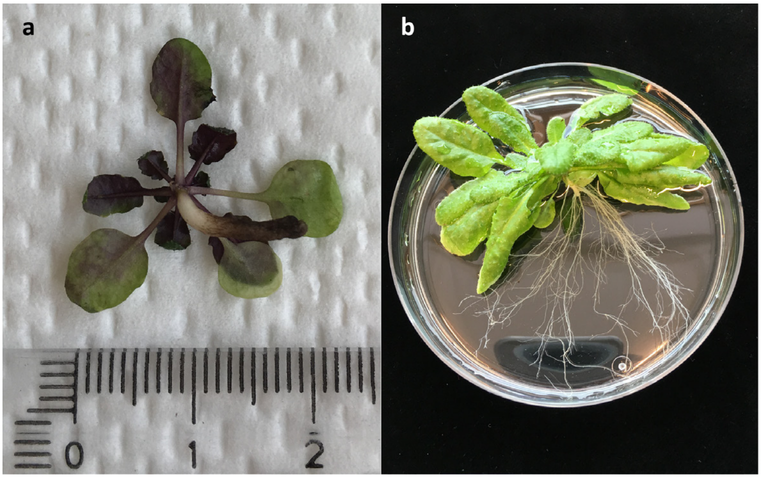


Figure 5. *Arabidopsis thaliana* Col-0 **a**. Plant displaying clubroot symptoms after six weeks in soil infested with *P. brassicae* e3 strain. **b**. Healthy plant after six weeks, submerged into a petri dish filled with water for easier observation of roots. Photo: Suzana Stjelja Arvelius.

Above- and below-ground symptoms of clubroot were obvious in comparison to a healthy *Arabidopsis* plant (**Figure 5b**). Furthermore, we observed phenotypic variation among infected plants. Some of the plants displayed stronger disease symptoms and their clubroots decomposed by time of harvesting. Thus, a careful monitoring of *Arabidopsis* plants infected with *P. brassicae* and observation of inter-plant variation as well as adjustment of harvesting time can help to reduce losses of the clubroot material for scientific use. This approach allowed us to collect two hundred *Arabidopsis* clubroots (**Figure 5a**) intended for RNA isolation and transcriptional analysis.

3.1.2 Spore isolation and DNA extraction

Based on numerous trials and various clubroot materials, we identified several key points in experimental procedures with *P. brassicae*. Selection of clubroots with right maturity and softness (not too firm nor decomposed) increases yield of resting spores while thorough removal of starch, debris and microbial contamination ensures spore purity (Paper I, Fig. 3 and 4). Next, meticulous grinding of frozen spores to a very fine powder helps to break thick spore walls and promotes extraction of nucleic acids. Other, more general precautions include avoiding contamination by carry-over phenol and chloroform and organic compounds after DNA extraction based on a CTAB (cetyltrimethylammonium bromide) protocol. For DNA precipitation with ethanol, a high salt molarity needs to be lowered for samples that contain less polysaccharides. Otherwise, large amounts of salt can precipitate and reduce DNA yield. These key points provided robustness of the protocol and enabled multiple extractions of *P. brassicae* DNA, with a high concentration and quality that fulfilled stringent requirements for PacBio sequencing (Paper I, Fig. 5). The protocol was further tested by extraction of total DNA from roots and clubroots intended for PacBio long-amplicon metagenomic sequencing.

3.2 PacBio-refined genomes of *P. brassicae* e3 strain

3.2.1 Nuclear genome assembly (Paper II)

PacBio RS II sequencing and *de-novo* assembly generated 20 contigs that encompassed a total size of 25.25 Mb (Paper II, Fig. 1). The assembly contiguity was therefore substantially improved in comparison with the first genome draft based on Illumina/454 sequences (Schwelm et al., 2015). Furthermore, long-read PacBio data enabled resolution of repeat-rich telomeric and sub-telomeric regions and large complex regions.

Telomeres were identified on 18 contigs of which 13 were assembled telomere-to-telomere and 5 contigs terminated with a single telomere (Paper II, Fig. 1). Telomeric sequences could not be assembled for the two remaining contigs. Since not all contigs were fully resolved chromosomes and because completeness of genomic sequences is generally difficult to claim, the “chromosome-level” assembly will be used to make this distinction.

The exact number and size of *P. brassicae* chromosomes is not yet known and 5 to 20 haploid chromosomes have been suggested (**chapter 1.2.5**). The number of the PacBio contigs corresponds to 20 haploid chromosomes that were indicated by a karyotype analysis based on counting of synaptonemal complexes (SC) in *P. brassicae* pachytene nuclei (Braselton, 1982). Despite the tempting conclusiveness, some of incomplete PacBio contigs may belong to a single chromosome and/or yet unresolved chromosome(s) might exist. This leads to a question on how large portion of the e3 nuclear genome remains to be recovered? About one quarter of a Mb might be an answer, according to a total genome size of 25.5 Mb that was estimated based on *k*-mer analysis of Illumina reads in the first genome draft (Schwelm et al., 2015). However, this estimate is dependent on the sequencing data and should be carefully interpreted when directly compared with the PacBio assembly. Sizes of all genome assemblies which are currently available for multiple strains of *P. brassicae* are within a similar range (24.04 - 25.25 Mb) while *S. subterranea* and *P. betae* have somewhat larger assemblies (28.08 and 27.08 Mb, respectively) (**Table 1**). Despite having the largest size, the PacBio assembly contains the lowest number of predicted protein-coding genes (9,231) among *P. brassicae* annotated assemblies (**chapter 1.2.5**). Besides genetic variation between strains, a lower level of gene fragmentation and a better resolution of repeat-rich regions in long-read assemblies (Oren et al., 2016; Paajanen et al., 2019) may explain those differences.

3.2.2 Repetitive sequences are more abundant (Paper II, III)

The content of repetitive sequences in *P. brassicae* genome assemblies is described as low (≤ 2.0%) based on information from two Canadian and a Chinese strain (Rolfe et al., 2016; Bi et al., 2019). In comparison, a significantly higher content (5.4%) was reported for the European e3 strain (Schwelm et al., 2015). These results may reflect genetic differences, particularly since chromosome-length polymorphism among strains has been proposed based on hybridization patterns of repetitive fragments in Southern analysis (Graf et al., 2004). In addition, similar sequencing platforms were used to generate those genome assemblies, reducing the possibility of sequence variation due to technical reasons (**Table 1**).

Our analysis of the PacBio data yielded 11.5% of repetitive sequences. This represents an increase of ~113% for the e3 strain in comparison to its Illumina/454 assembly (Schwelm et al., 2015). *De-novo* annotation revealed that among 2.9 Mb of repetitive sequences, interspersed repeats were by far the most abundant and included several known families of TEs and unclassified repeats (Paper III, Table 1). These results indicate that repetitive sequences are more abundant in the e3 nuclear genome than previously estimated. To investigate whether this is a common feature for *P. brassicae*, (re)sequencing of multiple strains by technologies that can improve resolution of repeat-rich regions is needed.

We identified telomeric sequences as a simple TTTTAGGG (T_4AG_3) repeat, spanning the first 200 to 300 bp at the contig ends (Paper II). The following 10 to 40 kb were encompassed by sub-telomeric regions, rich in the interspersed repeats. Next, large regions with numerous *Gypsy* and *Copia* retrotransposons and unclassified repeats were found along the contigs. These repeat-rich islands displayed extensive structural variation including duplications, palindromes and mini- and microsatellites, visualized by contig dot-plot self-similarity comparisons (Paper III, Fig. 1). Signatures of multiple genomic rearrangements were further detected in these regions by analysis of inter-chromosomal synteny (Paper III, Fig. 2). In total, we characterized 25 complex regions, at least one on each of the 20 nuclear contigs.

Whether the complex regions may represent putative centromeres of *P. brassicae* chromosomes remains to be further explored. This assumption is based on their structural complexity, repetitive nature, lower gene content and presence on each of the 20 nuclear contigs. However, large variation in structure and size is generally present among centromeres (Talbert and Henikoff, 2020) and characterization of centromere elements often requires a combination of imaging, experimental and sequencing data (Yadav et al., 2019; Sankaranarayanan et al., 2020). Divergent centromere elements are reported for organisms in the SAR group, from short and extremely AT-rich centromeres in *Plasmodium falciparum* (Vembar et al., 2016; Verma and Surolia, 2018) to larger and more complex centromeres in *Phytophthora sojae* (Fang et al., 2020). We found several genes coding for centromere-associated and chromosome segregation proteins in the PacBio assembly. However, their genomic location as well as location of shorter AT-rich sequences were not limited to the complex regions.

A deeper analysis of the *P. brassicae* annotation data identified nearly 190 repeat families (a fundamental unit of repeat and TE classification). More than a half of these families were found across the 25 complex regions, displaying a mosaic distribution. The most abundant in the PacBio assembly was family-6, annotated in total on 180 kb and distributed on 17 contigs. With a 5 kb long consensus sequence, family-6 is an interspersed repeat that lacks classification information. Interestingly, we found that the *P. brassicae*-specific repetitive element H4 (Klewer et al., 2001; Graf et al., 2001) is a part of the family-6 sequence. The H4 element still lacks similarity with any of the sequences deposited in the NCBI non-redundant nucleotide database (BLAST searches, October 2021) and our re-annotation of family-6 yielded inconclusive results.

The H4 element was used for characterization of chromosome-length polymorphism among *P. brassicae* strains (Graf et al., 2004). Furthermore, based on the H4 hybridization patterns chromosomal rearrangements were proposed to occur within a single isolate and without sexual recombination (Fähling et al., 2004). It is not yet clear whether sexual recombination occurs in *P. brassicae* (**chapter 1.2.2**). We identified several transcriptionally active SC-associated and meiosis-related orthologs in the PacBio assembly (Paper II). SC is a protein structure that connects homologous chromosomes and promotes recombination (Hesse et al., 2019). Twenty SCs were identified by electron microscopy in *P. brassicae* pachytene nuclei (Braselton, 1982). These findings in combination with the SC orthologs support occurrence of genetic recombination events in *P. brassicae*. However, it cannot be excluded that the *P. brassicae* orthologs may have other function(s) since sequence variation is a common feature among SC orthologs.

The analysis of repetitive sequences in *P. brassicae* leads to several interesting questions. What role(s) may the complex regions have in the genome organization? Are repetitive sequences involved in chromosomal rearrangements and recombination events? Is there a link between genome distribution of TEs and unclassified repeats and variation of virulence? Analysis of effector candidates (**chapter 3.2.4**) might provide initial clues to this question. Next, is there a mechanism behind the family-6 wide genome distribution? Can functional role(s) be assigned to the family-6 sequence?

3.2.3 Effector candidates (Paper II, III)

Based on the predicted nuclear gene models (9,231) and their amino acid sequences (Paper II) we selected 314 small, secreted proteins (SSPs) as effector candidates of *P. brassicae* (Paper III). We aimed at utilizing a larger number of potential effectors to investigate their genome distribution and motif enrichment. Therefore, we did not apply filtering by gene expression levels, used for selection of effector candidates (92) in the Illumina/454 genome assembly (Schwelm et al., 2015). Furthermore, we included proteins with a low cysteine content since only few cysteine-rich proteins were found among *P. brassicae* putative effectors, highly expressed during host infection (Pérez-López et al., 2020). Among the 314 SSPs, 204 were classified as cysteine-rich proteins (≥4 cysteine residues).

Following the general trend among *P. brassicae* proteins, functional annotation was available for about a half of the PacBio predicted effectors. Among these, proteins with leucine rich repeats, ankyrin repeats, kinase and RING (Really Interesting New Gene) domains were the most abundant. Next, serine aminopeptidases and lipases and glycosyl hydrolases were found as well as previously reported effectors including methyltransferases (Ludwig-Müller et al., 2015) and cysteine-rich and chitin recognition proteins, kazal-type serine and papain cysteine proteases, polysaccharide deacetylases and fasciclin and thaumatin proteins (Schwelm et al., 2015).

Gene expression values for the 314 SSPs are based on RNA-seq data (Schwelm et al., 2015), mapped to the PacBio genome assembly (Paper II). Even though only descriptive, these data can provide valuable information from three life-stages of *P. brassicae*, including mature and germinating resting spores and plasmodia. According to these values, a number of the effector candidates was among the most highly expressed genes (5-7%) in the *P. brassicae* genome.

Since *P. brassicae* is a true intracellular pathogen, it is likely that a large majority of its effectors target plant subcellular compartments. About 75% of the 314 SSPs were classified as cytoplasmic effectors by ApoplastP, a tool for prediction of effector plant subcellular localization (Sperschneider et al., 2018). Localizer (Sperschneider et al., 2017) further predicted which of the effector candidates can solely target a single compartment (plant nucleus, chloroplast or mitochondria) or use these compartments as multiple targets.

The commonly applied size criterion (≤400 aa) is an arbitrary threshold for selection of effector candidates and larger size effectors are considered in prediction pipelines (Lo Presti et al., 2015; Sperschneider et al., 2015). Thus, *P. brassicae* proteins with more than 400 aa and a secretion signal and absence of transmembrane domains may have a potential effector function. We identified 190 such proteins in the PacBio assembly. A NUDIX hydrolase was recently highlighted as an important pathogenicity factor for the clubroot development (Daval et al., 2020). While no NUDIX domain was annotated among the 314 SSPs, a NUDIX phosphohydrolase was found among the 190 proteins. This set of proteins, in combination with other lines of evidence, may serve as an additional resource for search of *P. brassicae* candidate effectors.

3.2.4 TEs in vicinity of effector candidates (Paper III)

For the first time, we can glance over the genome landscape surrounding *P. brassicae* effector candidates. Thanks to the "chromosome-level" assembly and annotation of repetitive sequences we were able to explore genome distribution of the 314 predicted effectors.

The effector candidates were found on all *P. brassicae* contigs, dispersed from the sub-telomeric regions and further along the contigs. They were located in gene-dense as well as in gene-poor repetitive sequences, including the complex regions on few contigs. It seems therefore that strong signatures of compartmentalization are not apparent in the *P. brassicae* nuclear genome. About 30% of the predicted effectors were harbored by four chromosomes (Paper III, Table 2; Fig. 4a). To explore whether such distribution represents a prediction bias or is somehow linked to the genome organization, chromosome-level assemblies from other *P. brassicae* strains and pathotypes are necessary.

Our analysis showed that at least one TE or an unclassified interspersed repeat was located within a 3 kb distance from more than 37% of the effector candidates. These effector genes were often annotated on the opposite strand from their neighboring repeat elements, sometimes with overlapping positions. Interestingly, group II introns were found in the *P. brassicae* mitochondrial genome, overlapping exon-intron boundaries in protein-coding genes (Paper II). The distribution patterns of TEs and group II introns might be simply a reflection of the compact nature of *P. brassicae* genomes. On the other hand, these patterns may provide a hint towards strategies involved in splicing mechanisms and gene expression regulation. Clusters of effector genes that co-occur in close proximity are suggested to facilitate coordinated expression and quick modifications (Palmer and Keller, 2010; Frantzeskakis et al., 2018). Clusters with two to three effector candidates (consecutive or separated by one, other gene) were found on majority of the *P. brassicae* contigs. Some of these clusters harbored genes within the same functional group (e.g. polysaccharide deacetylases). Furthermore, various TEs such as *Copia* and *Gypsy* retrotransposons and unclassified repeats, including family-6 were found embedded within the clusters (Paper III, Fig. 4b). Further analyses are needed to decipher whether such organization has a role in effector sequence diversification and regulation of gene expression in the clubroot pathogen.

3.2.5 Motif enrichment of effector candidates (Paper III)

Only few *P. brassicae* effector candidates are reported to contain host-targeting motifs that are frequently found in fungal and oomycete effectors (RxLR, LxFLAK, CHXC, DEER and Pexel) (Schwelm et al., 2015, Rolfe et al., 2016; Pérez-López et al., 2020). To utilize the entire set of 314 effector candidates, we searched for a motif enrichment rather than looking for occurrence of individual motifs. About 60% of the predicted effectors were enriched in a motif rich in leucine, valine and alanine amino acids. We are currently performing analyses to gain understanding of potential functional role(s) of the motif.

3.2.6 Mitochondrial genome (Paper II)

When this PhD project started, no complete mitochondrial (mt) sequence of the *P. brassicae* e3 strain was available. This was one of the main reasons for application of the long-read sequencing.

Thanks to resolution of a 12 kb repeat-rich region, a single mt contig (114.6 kb) was generated in the new assembly. The region was not found in mt sequences of three other *P. brassicae* strains (Paper II, Fig. 2) nor in the closely related plasmodiophorid, *S. subterranea* (Paper II, Fig. 4). Since the region was also absent in the incomplete mt sequence of e3, based on Illumina/454 assembly, observed sequence variation was most likely caused by technical reasons. Tandem mini- and microsatellite repeats and palindromic AT-rich sequences, characterized by large variation in GC content were found in the 12 kb region.

A circular structure of the mt genome was confirmed by identification of overlapping sequences. Prediction of gene models was more challenging than anticipated. Extensive annotation revealed an intron-rich mt genome with a compact and complex organization (Paper II, Fig. 3). Introns were found in rRNA genes (small and large ribosomal subunit) and particularly in protein-coding genes. For example, 13 introns were annotated within the *cox*1 gene. Defining intron-exon boundaries by using common splicing sites caused fragmentation of coding sequences. Such sequence interruptions were corrected by application of atypical splicing sites, indicating that an alternative splicing mechanism is used in the mt genome of *P. brassicae*. Further analysis identified several intron-encoded proteins and group II introns that overlap intron-exon boundaries (Paper II, Fig. 3).

Together these results show complex intron patterns in the mt genome and lead to several questions. Are intron splicing events part of a *P. brassicae* regulatory network? Can similar patterns be observed in the nuclear genome as well? There is no information on gene regulation in the clubroot pathogen. Deletion of introns in yeast *cob* and *cox*1 mt genes (components in the electron transport chain) showed that inefficient splicing is not detrimental but instead necessary for the normal life (Rudan et al., 2018). Higher transcript levels (present after intron removal) were perceived too stressful, resulting in negative effects on yeast growth and lifespan. Different life-stages of *P. brassicae* can require activity changes and it cannot be ruled out that intron splicing events may be involved in regulation of such changes.

Numerous introns and the 12 kb repetitive region as well as variation in intergenic regions contributed to a large size difference between the mt genomes of *P. brassicae* (114.6 kb) and *S. subterranea* (37.7 kb). Despite these differences, two plasmodiophorids shared high sequence synteny (Paper II, Fig. 4) and similar gene content and gene order (Paper II, Fig. 3). Their close relationship was further confirmed by phylogenetic analyses based on 12 mt protein-coding genes from 67 organisms (Paper II, Fig. 5). The plasmodiophorids clustered together with two chlorarachniophyte algae and a cercomonad flagellate, representing rare Rhizaria with available mt genome sequences.

3.3 Remaining work

A number of sequence and experimental analyses remain to be performed to:

- Further elucidate importance and functional roles of the predicted effector candidates,
- Investigate sequence similarities to animal parasites and algae,
- Re-annotate the family-6 repeat sequence and gain understanding about its possible role(s) in the nuclear genome,
- Investigate presence and distribution of centromere-specific elements along the nuclear contigs.

4. Conclusions

While working with the *P. brassicae* e3 strain and its genomic data I encountered various (un)expected challenges which shaped our analysis and often led to new insights.

The main conclusions from this thesis are:

- *P. brassicae* DNA with quality and quantity suitable for genome sequencing and other molecular analysis can be obtained by application of our protocol.

- The long-read *de-novo* assembly has significantly improved contiguity thanks to resolved highly repetitive regions and closed sequence gaps.

- The nuclear genome (25.2 Mb) is represented by 20 contigs. For the first time, telomeric sequences (T_4AG_3) are identified and 13 contigs are assembled from telomere-to-telomere while 5 contigs terminate with a single telomere. Telomeres could not be assembled for the two remaining contigs.

- Repetitive sequences represent 11.5% of the genome assembly. Interspersed repeats, including TEs and unclassified repeats are the most abundant group. They are particularly clustered in the sub-telomeric regions and the complex regions with extensive structural variation, found on each contig. The most abundant repetitive sequence in the genome is family-6, found on 17 contigs and annotated as an unclassified repeat.

- The nuclear genome has 9,231 predicted protein-coding genes, of which more than 40% lack functional annotation.

- Small, secreted nuclear proteins are selected as effector candidates, including cysteine-rich (204) and cysteine-low (110) proteins. They are distributed along all contigs and clusters with two to three effector genes are often observed. TEs and unclassified interspersed repeats are found within a 3 kb distance from 37% of the predicted effectors. Enrichment with a motif rich in valine, leucine and alanine amino acids is detected among 60% of the effector candidates.

- The mt genome (114.6 kb) has a circular sequence which contains a 12 kb long, previously unresolved repetitive region. Numerous intron-rich genes contribute to the mt genome compact organization. Many of these genes have a complex intron pattern (with intron-encoded proteins) and an alternative splicing mechanism (with atypical acceptor-donor sites). Furthermore, group II introns are found to overlap intron-exon boundaries.

- The close relationship between *P. brassicae* and *S. subterranea* is confirmed by comparison of their mt genomes (sequence synteny, gene content and gene order) and by phylogenetic analysis based on mt proteins.

These findings together with the long-read genome sequences create a valuable platform for comparative, prediction-based and functional studies of *P. brassicae* and its interaction with a plant host.

5. Future perspectives

A “wish list” of methods and resources that are essential for taking the understanding of *P. brassicae* and its plant interactions forward is quite long. Highly prioritized are projects such as development of heterologous expression system(s) for evaluation of *P. brassicae* gene function, a protocol for genetic transformation, clarification of the life-events and existence of the sexual phase, identification of the molecular basis for virulence variation, DNA marker tests for pathotyping with a high differentiating capacity and identification of plant host targets and the molecular basis of resistance.

The work in this thesis has revealed unknown features in the *P. brassicae* genomes and a number of knowledge gaps. To further advance, there is a need for:

- High-quality genomes from *P. brassicae* multiple strains and pathotypes to help deciphering strain genetic variation.

- Genome sequences from phagomyxids causing galls on brown algae and seagrasses to help resolving the complex obligate biotrophic life-stages and interactions with a host.

- Incorporating knowledge and utilizing sequence resources from galling insects and animal parasites to help understanding the infection biology.

- Genome sequences from Rhizaria and other protists to help resolving the evolutionary history of *P. brassicae* and possible connections including bacteria-algae-plant-*P. brassicae*.

References

Aist, J.R. and Williams, P.H. 1971.The cytology and kinetics of cabbage root hair penetration by *Plasmodiophora brassicae*. *Can. J. Bot.* 49, 2023-2034.

Ayers, G.W. 1957. Races of *Plasmodiophora brassicae*. *Can. J. Bot.* 35, 923-932.

Balotf, S., Tegg, R.S., Nichols, D.S. and Wilson, C.R. 2021. Spore germination of the obligate biotroph *Spongospora subterranea*: transcriptome analysis reveals germination associated genes. *Front. Microb.* 12, 691877.

Bi, K., He, Z., Gao, Z. et al. 2016. Integrated omics study of lipid droplets from *Plasmodiophora brassicae*. *Sci. Rep.* 6, 36965.

Białas A., Zess E.K., De la Concepcion J.C. et al. 2018. Lessons in effector and NLR biology of plant-microbe systems. *Mol. Plant-Microbe Interact.* 31, 34-45.

Biard, T., Krause J.W., Stukel M.R. and Ohman, M.D. 2018. The significance of giant phaeodarians (Rhizaria) to biogenic silica export in the California current ecosystem. *Gl. Biogeochem. Cycl.* 32, 987-1004.

Biard, T., Stemmann, L., Picheral, M. et al. 2016. *In situ* imaging reveals the biomass of giant protists in the global ocean. *Nature* 532, 504-507.

Botero, A., García, C., Gossen, B.D. et al. 2019. Clubroot disease in Latin America: distribution and management strategies. *Plant Pathol.* 68, 827-833.

Braselton, J.P. 1982. Karyotypic analysis of *Plasmodiophora brassicae* based on serial thin sections of pachytene nuclei. *Can. J. Bot.* 60, 403-408.

Braselton, J.P. 1995. Current status of the plasmodiophorids. *Crit. Rev. Microb.* 21, 263-275.

Bryan R.J., Trese A.T. and Braselton, J.P. 1996. Molecular karyotypes for the obligate, intracellular, plant pathogens, *Plasmodiophora brassicae* and *Spongospora subterranea*. *Mycologia* 88, 358-360.

Buczacki, S.T., Toxopeus, H., Mattusch, P. et al. 1975. Study of physiological specialization in *Plasmodiophora brassicae*: proposals for attempted rationalisation through an international approach. *Trans. Br. Mycol. Soc.* 65, 295-303.

Bulman S. and Neuhauser S. 2016. Phytomyxea. In: Handbook of the Protists (Archibald, J. et al. eds). *Springer, Cham*. pp. 1-21.

Bulman S.R., Kuhn S.F., Marshall J.W. and Schnepf, E. 2001. A phylogenetic analysis of the SSU rRNA from members of the Plasmodiophorida and Phagomyxida. *Protist* 152, 43-51.

Bulman, S. and Braselton, J.P. 2014. Rhizaria: Phytomyxea. In: The Mycota VII, Part A, Systematics and Evolution (McLaughlin, D.J. and Spatafora, J.W. eds). Springer Berlin Heidelberg pp. 99-112.

Burki F. and Pawlowski, J. 2006. Monophyly of Rhizaria and multigene phylogeny of unicellular bikonts. *Mol. Biol. Evol.* 23, 1922-1930.

Burki F., **Shalchian-Tabrizi K., Minge M. et al.** 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS ONE* 2, e790.

Burki, F. and Keeling, P.J. 2014. Rhizaria. *Curr. Biol.* 24, R103-R107.

Burki F. 2014. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* 6, a016147.

Burki F., Roger A.J., Brown M.W. and Simpson A.G.B. 2020. The new tree of eukaryotes. *Trends. Ecol. Evol.* 35, 43-55.

Cao T., Manolii V.P., Zhou Q. et al. 2019. Effect of canola (*Brassica napus*) cultivar rotation on *Plasmodiophora brassicae* pathotype composition. *Can. J. Plant Sci.* 100, 218-225.

Cavalier-Smith T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int. J. Syst. Evol. Microbiol.* 52, 297-354.

Cavalier-Smith, T., Caho, E.E., and Lewis, R. 2018. Multigene phylogeny and cell evolution of chromist infrakingdom Rhizaria: contrasting cell organisation of sister phyla Cercozoa and Retaria. *Protoplasma* 255, 1517-1574.

Chai, A.L., Li, J.P., Xie, X.W., Shi, Y.X. and Li, B.J. 2016. Dissemination of *Plasmodiophora brassicae* in livestock manure detected by qPCR. *Plant Pathol.* 65, 137-144.

Chen, W., Li, Y., Yan, R. et al. 2021. SnRK1.1-mediated resistance of *Arabidopsis thaliana* to clubroot disease is inhibited by the novel *Plasmodiophora brassicae* effector PBZF1. *Mol. Plant Pathol.* 22, 1057-1069.

Cheng S., Melkonian M., Smith S.A. et al. 2018. 10KP: A phylodiverse genome sequencing plan. *GigaScience* 7, giy013.

Ciaghi, S., Neuhauser, S. and Schwelm, A. 2018. Draft genome resource for the potato powdery scab pathogen *Spongospora subterranea*. *Mol. Plant-Microbe Interact.* 31, 1227-1229.

Ciaghi, S., Schwelm, A. and Neuhauser, S. 2019. Transcriptomic response in symptomless roots of clubroot infected kohlrabi (*Brassica oleracea* var. *gongylodes*) mirrors resistant plants. *BMC Plant Biol.* 19, 288.

Datnoff, L.E., Kroll, T.K. and Lacy, G.H. 1987. Efficacy of chlorine for decontaminating water infested with resting spores of *Plasmodiophora brassicae. Plant Dis.* 71, 734-736.

Daval, S., Belcour, A., Gazengel, K. et al. 2019. Computational analysis of the *Plasmodiophora brassicae* genome: mitochondrial sequence description and metabolic pathway database design. *Genomics* 111, 1629-1640.

Daval, S., Gazengel, K., Belcour, A. et al. 2020. Soil Microbiota influences clubroot disease by modulating *Plasmodiophora brassicae* and *Brassica napus* transcriptomes. *Microb. Biotechnol.* 13, 1648-1672.

Decroës, A., Calusinska, M., Delfosse, P. et al. 2019. First draft genome sequence of a *Polymyxa* genus member, *Polymyxa betae*, the protist vector of rhizomania. *Microb. resour. announ*. 8, e01509-18.

del Campo, J., Sieracki M.E, Molestina, R. et al. 2014. The others: our biased perspective of eukaryotic genomes. *Trends Ecol. Evol*. 29, 252-259.

Diederichsen, E., Frauen, M., Linders, E.G.A. et al. 2009. Status and perspectives of clubroot resistance breeding in crucifer crops. *J. Plant Growth Regul.* 28, 265-281.

Dixon, G.R. 2009a. The occurrence and economic impact of *Plasmodiophora brassicae* and clubroot disease. *J. Plant Growth Regul.* 28, 194-202.

Dixon, G.R. 2009b. *Plasmodiophora brassicae* in its environment. *J. Plant Growth Regul.* 28, 212-228.

Dixon, G.R. 2014. Clubroot (*Plasmodiophora brassicae* Woronin) - an agricultural and biological challenge worldwide. *Can. J. Plant Pathol*. 36, 5-8.

Dixon, G.R. 2017. Managing clubroot disease (caused by *Plasmodiophora brassicae* Wor.) by exploiting the interactions between calcium cyanamide fertilizer and soil microorganisms. *J. Agric. Sci.* 155, 527-543.

Dodds, P.N. and Rathjen, J.P. 2010. Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat. Rev. Genet.* 11, 539-548.

Dong, S., Raffaele, S. and Kamoun, S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. *Curr. Opin Genet. Develop*. 35, 57-65.

Dylewski, D.P., Braselton, J.P. and Miller, C.E. 1978. Cruciform nuclear division in *Sorosphaera veronicae*. *Amer. J. Bot.* 65, 258-267.

Elliott, J.K., Simpson, H., Teesdale, A. et al. 2019. A novel phagomyxid parasite produces sporangia in root hair galls of eelgrass (*Zostera marina*). *Protist* 170,64-81.

Fähling, M., Graf, H. and Siemens, J. 2004. Characterization of a single-spore isolate population of *Plasmodiophora brassicae* resulting from a single club. *J. Phytopathol*. 152, 438-444.

Falloon, R.E., Merz, U., Butler, R.C. et al. 2016. Root infection of potato by *Spongospora subterranea*: knowledge review and evidence for decreased plant productivity. *Plant Pathol.* 65, 422-434.

Fang, Y., Coelho M.A., Shu, H. et al. 2020. Long transposon-rich centromeres in an oomycete reveal divergence of centromere features in Stramenopila-Alveolata-Rhizaria lineages. *PLoS Genet* 16, e1008646.

Fiore-Donno, A.M, Richter-Heitmann, T., Degrune, F. 2019. Functional traits and spatio-temporal structure of a major group of soil protists (Rhizaria: Cercozoa) in a Temperate Grassland. *Front. Microbiol.* 10, 1332.

Fiore-Donno, A.M., Richter-Heitmann, T. and Bonkowski, M. 2020. Contrasting responses of protistan plant parasites and phagotrophs to ecosystems, land management and soil properties. *Front. Microbiol.* 11, 1823.

Frantzeskakis, L., Kracher, B., Kusch, S. et al. 2018. Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. *BMC Genomics.* 19, 381.

Frantzeskakis, L., Kusch, S. and Panstruga, R. 2019. The need for speed: compartmentalized genome evolution in filamentous phytopathogens. *Mol. Plant Pathol.* 20, 3-7.

Fu, H., Yang, Y., Mishra, V. et al. 2020. Most *Plasmodiophora brassicae* populations in single canola root calls from Alberta fields are mixtures of multiple strains. *Plant. Dis*. 104, 116-120.

Fuchs, H. and Sacristan, M.D. 1996. Identification of a gene in *Arabidopsis thaliana* controlling resistance to clubroot (*Plasmodiophora brassicae*) and characterization of the resistance response. *Mol. Plant-Microbe Interact.* 9, 91-97.

Gout, L., Fudal, I., Kuhn, M.L. et al. 2006. Lost in the middle of nowhere: the *AvrLm1* avirulence gene of the Dothideomycete *Leptosphaeria maculans*. *Mol. Microbiol*. 60, 67-80.

Graf, H., Sokolowski, F., Klewer, A. et al. 2001. Electrophoretic karyotype of the obligate biotrophic parasite *Plasmodiophora brassicae* Wor. *J. Phytopathol*. 149, 313-318.

Graf, H., Fähling, M. and Siemens, J. 2004. Chromosome polymorphism of the obligate biotrophic parasite *Plasmodiophora brassicae. J. Phytopathol*. 152, 86-91.

Grattepanche, J.D., Walker, L.M., Ott, B.M. et al. 2018. Microbial diversity in the eukaryotic SAR clade: illuminating the darkness between morphology and molecular data. *Bioessays* 40: e1700198.

Gravot, A., Lemarié S., Richard, G. et al. 2016. Flooding affects the development of *Plasmodiophora brassicae* in *Arabidopsis* roots during the secondary phase of infection. *Plant Pathol*. 65, 1153-1160.

Groussin, M., Pawlowski, J. and Yang, Z. 2011. Bayesian relaxed clock estimation of divergence times in foraminifera. *Mol. Phylogenet. Evol*. 61, 157-166.

Guidi, L., Chaffron, S., Bittner, L. et al. 2016. Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 532, 465-470.

Gutiérrez, P., Bulman, S., Alzate, J. et al. 2016. Mitochondrial genome sequence of the potato powdery scab pathogen *Spongospora subterranea*. *Mitochon. DNA Part A* 27, 58-59.

Hatakeyama, K., Fujimura, M., Ishida, M. and Suzuki, T. 2004. New classification method for *Plasmodiophora brassicae* field isolates in Japan based on resistance of F1 cultivars of Chinese cabbage (*Brassica rapa* L.) to clubroot. *Breeding Sci*. 54, 197-201.

Heo, S.H, Jang, S.J, Choi, J.S. et al. 2009. Chinese cabbage clubroot pathogen, *Plasmodiophora brassicae*, is genetically stable. *Mycobiology* 37, 225-229.

Hesse, S., Zelkowski, M., Mikhailovam, E.I. et al. 2019. Ultrastructure and dynamics of synaptonemal complex components during meiotic pairing and synapsis of standard (A) and accessory (B) rye chromosomes. *Front. Plant Sci*. 10, 773.

Hittorf, M., Letch-Pramarer, S., Windegger, A. et al. 2020. Revised taxonomy and expanded biodiversity of the Phytomyxea (Rhizaria, Endomyxa). *J. Euk. Microbiol.* 67, 648-659.

Holtz, M.D., Hwang, S-F. and Strelkov, S.E. 2018. Genotyping of *Plasmodiophora brassicae* reveals the presence of distinct populations. *BMC Genomics* 19, 254.

Holtz, M.D., Hwang, S-F., Manolii, V.P. et al. 2021a. Development of molecular markers to identify distinct populations of *Plasmodiophora brassicae*. *Eur. J. Plant Pathol.* 159, 637–654.

Holtz, M.D., Hwang, S-F., Manolii, V.P. and Strelkov, S.E. 2021b. Molecular evaluation of *Plasmodiophora brassicae* collections for the presence of divergent genetic pathogen populations before and after the release of clubroot resistant canola. *Can. J. Plant Pathol.*

Howard, R.J., Strelkov, S.E. and Harding, M.W. 2010. Clubroot of cruciferous crops – new perspectives on an old disease. *Can. J. Plant Pathol*. 32, 43-57.

Ingram, D.S. and Tommerup, I.C. 1972. The life history of *Plasmodiophora brassicae* Woron. *Proc. R. Soc. Lond*. B, 180, 103–112.

Irani, S., Trost, B., Waldner, M. et al. 2018. Transcriptome analysis of response to *Plasmodiophora brassicae* infection in the Arabidopsis shoot and root. *BMC Genomics* 19, 23.

Ito, S.I., Yano, S., Tanaka, S. and Kameya-Iwaki, M. 1994. The use of resting spore spheroplasts in the DNA analysis of *Plasmodiophora brassicae*. *Ann. Phytopathol. Soc. Jap*. 60, 491-495.

Jaswal, R., Kiran, K., Rajarammohan, S. et al. 2020. Effector biology of biotrophic plant fungal pathogens: current advances and future prospects. *Microb. Res*. 241, 26567.

Jeong, J-Y., Robin, A.H.K., Natarajan, S. et al. 2018. Race- and isolate-specific molecular marker development through genome realignment enables detection of Korean *Plasmodiophora brassicae* isolates, causal agent of clubroot disease. *Plant Pathol. J.* 34, 506-513.

Jones, J.D. and Dangl, J.L. 2006. The plant immune system. *Nature* 444, 323-329.

Kageyama, K. and Asano, T. 2009. Life cycle of *Plasmodiophora brassicae*. *J. Plant Growth Regul.* 28, 203-211.

Kamoun, S. 2006. A catalogue of the effector secretome of plant pathogenic oomycetes. *Annu. Rev. Phytopathol.* 44, 41-60.

Kanyuka, K., Ward, E. and Adams, M.J. 2003. *Polymyxa graminis* and the cereal viruses it transmits: a research challenge. *Mol. Plant Pathol.* 4, 393-406.

Karling, J.S. 1942. The Plasmodiophorales. New York, NY, USA.

Keeling, P.J, Burki, F., Wilcox, H.M. et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* 12, e1001889.

Keeling, P.J. and Burki, F. 2019. Progress towards the tree of eukaryotes. *Curr. Biol.* 29, R808-R817.

Kemen, E., Gardiner, A., Schultz-Larsen, T. et al. 2011. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*. *PLoS Biol.* 9, e1001094.

Kim, H., Jo, E.J., Choi, Y.H. et al. 2016. Pathotype classification of *Plasmodiophora brassicae* isolates using clubroot-resistant cultivars of Chinese Cabbage. *Plant pathol.* 32, 423-430.

Kitts, P.A., Church, D.M., Thibaud-Nissen, F. et al. 2016. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* 44, D73-80.

Klewer, A., Luerßen, H., Graf, H. and Siemens, J. 2001. Restriction fragment length polymorphism markers to characterize *Plasmodiophora brassicae* single-spore isolates with different virulence patterns. *J. Phytopath.* 149, 121-127.

Kolátková, V., Čepička, I., Gargiulo, G.M. et al. 2020. Enigmatic phytomyxid parasite of the alien seagrass *Halophila stipulacea*: new insights into its ecology, phylogeny, and distribution in the Mediterranean Sea. *Microb. Ecol.* 79, 631-643.

Kolátková, V., Čepička, I., Hoffman, R. and Vohník, M. 2021. *Marinomyxa* Gen. Nov. accommodates gall-forming parasites of the tropical to subtropical seagrass genus *Halophila* and constitutes a novel deep-branching lineage within phytomyxea (Rhizaria: Endomyxa). *Microb. Ecol.* 81, 673-686.

Krabberød, A.K., Orr, R.J.S., Bråte, J. et al. 2017. Single cell transcriptomics, mega-phylogeny, and the genetic basis of morphological innovations in Rhizaria. *Mol. Biol. Evol.* 34, 557-1573.

Kuginuki, Y., Yoshikawa, H., and Hirai, M. 1999. Variation in virulence of *Plasmodiophora brassicae* in Japan tested with clubroot-resistant cultivars of Chinese cabbage (*Brassica rapa* L. ssp. pekinensis). *Eur. J. Plant Pathol.* 105, 327-332.

LeBoldus, J.M, Manolii, V.P, Turkington, T.K. and Strelkov, S.E. 2012. Adaptation to Brassica host genotypes by a single-spore isolate and population of *Plasmodiophora brassicae* (clubroot). *Plant Dis*. 96, 833-838.

Lebreton, L., Guillerm-Erckelboudt, A.Y., Gazengel, K. et al. 2019. Temporal dynamics of bacterial and fungal communities during the infection of *Brassica rapa* roots by the protist *Plasmodiophora brassicae*. *PLOS ONE* 14, e0204195.

Liu, L., Qin, L., Zhou, Z. et al. 2020. Refining the life cycle of *Plasmodiophora brassicae. Phytopathol*. 110, 1704-1712.

Llopis Monferrer, N., Boltovskoy, D., Tréguer, P. et al. 2020. Estimating biogenic silica production of Rhizaria in the global ocean. *Global Biogeochem. Cycles* 34, 3.

Lo Presti, L., Lanver, D., Schweizer, G. et al. 2015. Fungal effectors and plant susceptibility. *Annu. Rev. Plant Biol*. 66, 513-45.

Ludwig-Müller, J., Prinsen, E., Rolfe, S.A. et al. 2009. Metabolism and plant hormone action during clubroot disease. *J. Plant Growth Regul*. 28, 229-244.

Ludwig-Müller, J., Jülke, S., Geiss, K. et al. 2015. A novel methyltransferase from the intracellular pathogen *Plasmodiophora brassicae* participates in methylation of salicylic acid. *Mol. Plant Pathol.* 16, 349-64.

MacFarlane, I. 1970. Germination of resting spores of *Plasmodiophora brassicae*. *Trans. Br. Mycol. Soc.* 55, 97-112.

Malinowski, R., Truman, W. and Blicharz, S. 2019. Genius architect or clever thief - how *Plasmodiophora brassicae* reprograms host development to establish a pathogen-oriented physiological sink. *Mol. Plant-Microbe Interact.* 32, 1259-1266.

Manzanares-Dauleux, M.J., Divaret, I., Baron, F. and Thomas, G. 2001. Assessment of biological and molecular variability between and within field isolates of *Plasmodiophora brassicae*. *Plant Pathol*. 50, 165-73.

McGrann, G.R.D., Grimmer, M.K., Mutasa-Göttgens, E.S. and Stevens, M. 2009. Progress towards the understanding and control of sugarbeet rhizomania disease. *Mol. Plant Pathol*. 10, 129-141.

McIlroy, D., Green, O.R. and Brasier, M.D. 2001. Paleobiology and evolution of the earliest agglutinated Foraminifera: platysolenites, spirosolenites and related forms. *Lethaia* 34, 13-29.

Miao, W., Song, L., Ba, S. et al. 2020. Protist 10,000 genomes project. *The Innovation* 1, 100058.

Murúa, P., Goecke, F., Westermeier, R. et al. 2017. *Maullinia braseltonii* sp. nov. (Rhizaria, Phytomyxea, Phagomyxida): a cyst-forming parasite of the bull kelp *Durvillaea* spp. (Stramenopila, Phaeophyceae, Fucales). *Protist* 168, 468-480.

Neuhauser, S. and Kirchmair, M. 2009. *Ligniera junci*, a plasmodiophorid rediscovered in roots of *Juncus* in Austria. *Osterr. Z. fur Pilzkd*. 18, 141-147.

Neuhauser, S., Bulman, S. and Kirchmair, M. 2010. Plasmodiophorids: the challenge to understand soil-borne, obligate biotrophs with a multiphasic life cycle. In: Current Advances in Molecular Identification of Fungi (Gherbawy, Y. and Voigt, K. eds). Springer Berlin Heidelberg pp. 51-78.

Neuhauser, S., Kirchmair, M. and Gleason, F.H. 2011. The ecological potentials of Phytomyxea ('plasmodiophorids') in aquatic food webs. *Hydrobiologia* 659, 23-35.

Neuhauser, S., Kirchmair, M., Bulman, S. et al. 2014. Cross-kingdom host shifts of phytomyxid parasites. *BMC Evol. Biol.* 14, 33.

Olszak, M., Truman, W., Stefanowicz, K. et al. 2019. Transcriptional profiling identifies critical steps of cell cycle reprogramming necessary for *Plasmodiophora brassicae*-driven gall Formation in Arabidopsis. *Plant J.* 97, 715-729.

Oren, M., Barela Hudgell, M.A., D'Allura, B. et al. 2016. Short tandem repeats, segmental duplications, gene deletion, and genomic instability in a rapidly diversified immune gene family. *BMC Genomics* 17, 900.

Paajanen, P., Kettleborough, G., López-Girona, E. et al. 2019. A critical comparison of technologies for a plant genome sequencing project. *GigaScience* 8, giy163.

Palmer, J.M. and Keller, N.P. 2010. Secondary metabolism in fungi: does chromosomal location matter?. *Curr. opin. microbiol.* 13, 431-436.

Pang, W., Liang Y., Zhan, Z. et al. 2020. Development of a Sinitic clubroot differential set for the pathotype classification of *Plasmodiophora brassicae*. *Front. Plant Sci.*11, 568771.

Parodi, E.R., Cáceres, E.J., Westermeier, R. and Müller, D.G. 2010. Secondary zoospores in the algal endoparasite *Maullinia ectocarpii* (Plasmodiophoromycota). *Biocell* 34, 45-52.

Pawlowski, J. 2000. Introduction to the molecular systematics of Foraminifera. *Micropaleontology* 46, 1-12.

Pérez-López, E., Hossain, M.M., Tu, J. et al. 2020. Transcriptome analysis identifies *Plasmodiophora brassicae* secondary infection effector candidates. *J. Eukaryot. Microbiol.* 67, 337-351.

Pérez-López, E., Hossain, M.M., Wei, Y. et al. 2021. A clubroot pathogen effector targets cruciferous cysteine proteases to suppress plant immunity. *Virulence* 12, 2327-2340.

Povolotskaya, I.S. and Kondrashov, F.A. 2010. Sequence space and the ongoing expansion of the protein universe. *Nature* 465, 922-926.

Raffaele, S. and Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10, 417-30.

Rennie, D.C., Holtz, M.D., Turkington, T.K. et al. 2015. Movement of *Plasmodiophora brassicae* resting spores in windblown dust. *Can. J. Plant Pathol.* 37, 188-196.

Rochon, D., Kakani, K., Robbins, M. and Reade, R. 2004. Molecular aspects of plant virus transmission by olpidium and plasmodiophorid vectors. *Annu. Rev. Phytopathol.* 42, 211-241.

Rolfe, S.A., Strelkov, S.E., Links, M.G. et al. 2016. The compact genome of the plant pathogen *Plasmodiophora brassicae* is adapted to intracellular interactions with host *Brassica* spp. *BMC Genom.* 17, 272.

Rudan, M., Bou Dib, P., Musa, M. et al. 2018. Normal mitochondrial function in *Saccharomyces cerevisiae* has become dependent on inefficient splicing. *eLife* 7, e35330.

Ruggiero, M., Gordon, D., Bailly, N. et al. 2015. A higher-level classification of all living organisms. *PLoS One* 10, e0119248.

Sánchez-Vallet, A., Fouché, S., Fudal, I. et al. 2018. The genome biology of dffector gene evolution in filamentous plant pathogens. Annu. Rev. Phytopathol. 56, 21-40.

Sankaranarayanan, SR., Ianiri, G., Coelho, M.A. et al. 2020. Loss of centromere function drives karyotype evolution in closely related *Malassezia* species. Elife 9, e53944.

Schnepf, E., Kühn, S.F. and Bulman, S. 2000. *Phagomyxa bellerocheae* sp. nov. and *Phagomyxa odontellae* sp. nov., Plasmodiophoromycetes feeding on marine diatoms. *Helgol. Mar. Res.* 54, 237-241

Schoch, C.L., Ciufo, S., Domrachev, M. et al. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020, baaa062.

Schuller, A. and Ludwig-Müller, J. 2016. Histological methods to detect the clubroot *Plasmodiophora brassicae* during its complex life cycle. *Plant Pathol.* 65, 1223-1237.

Schwelm, A., Fogelqvist, J., Knaust, A. et al. 2015. The *Plasmodiophora brassicae* genome reveals insights in its life cycle and ancestry of chitin synthases. Sci. Rep. 5, 11153.

Schwelm, A., Badstöber, J., Bulman, S. et al. 2018. Not in your usual Top 10: protists that infect plants and algae. Mol. Plant Pathol. 19, 029-1044.

Schwelm, A. and Ludwig-Müller, J. 2021. Molecular pathotyping of *Plasmodiophora brassicae*—genomes, marker genes, and obstacles. *Pathogens* 10, 259.

Sedaghatkish, A., Gossen, B.D., Yu, F. et al. 2019. Whole-genome DNA similarity and population structure of *Plasmodiophora brassicae* strains from Canada. *BMC Genom.*, 20, 744.

Shapiro-Ilan, D.I., Fuxa, J.R., Lacey, L.A. et al. 2005. Definitions of pathogenicity and virulence in invertebrate pathology. *J. Invertebr. Pathol.* 88, 1-7.

Sharma, K., Gossen, B.D. and McDonald, M.R. 2011. Effect of temperature on primary infection by *Plasmodiophora brassicae* and initiation of clubroot symptoms. *Plant Pathol.* 60, 830-8.

Sibbald, S. and Archibald, J. 2017. More protist genomes needed. *Nat. Ecol. Evol.* 1, 0145.

Sierra, R., Canas-Duarte, S. J., Burki, F. et al. 2016. Evolutionary origins of rhizarian parasites. *Mol. Biol. Evol.* 33, 980-983.

Singh, K., Tzelepis, G., Zouhar, M. et al. 2018. The immunophilin repertoire of *Plasmodiophora brassicae* and functional analysis of *PbCYP3* cyclophilin. *Mol. Genet. Genom.* 293, 381-390.

Somé, A., Manzanares, M.J., Laurens, F. et al. 1996. Variation for virulence on *Brassica napus* L. amongst *Plasmodiophora brassicae* collections from France and derived single-spore isolates. *Plant Pathol.* 45, 432-439.

Sperschneider, J., Dodds, P.N., Gardiner, D.M. et al. 2015. Advances and challenges in computational prediction of fffectors from plant pathogenic fungi. *PLoS Pathog.* 11, e1004806.

Sperschneider, J., Catanzariti, AM., DeBoer, K. et al. 2017. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. *Sci. Rep.* 7, 44598.

Sperschneider, J., Dodds, P.N., Singh, K.B. and Taylor, J.M. 2018. ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytol.* 217, 1764-1778.

Stergiopoulos, I. and de Wit, P.J.G.M. 2009. Fungal Effector Proteins. *Annu. Rev. Phytopathol.* 47, 233-263.

Stjelja, S., Fogelqvist, J., Tellgren-Roth, C. and Dixelius, C. 2019. The architecture of the *Plasmodiophora brassicae* nuclear and mitochondrial genomes. *Sci. Rep.* 9, 15753.

Strassert, J.F.H., Mahwash, J., Mylnikov, A.P. et al. 2019. New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Mol. Biol. Evol.* 36, 757-765.

Strehlow, B., de Mol, F. and Struck, C. 2014. History of oilseed rape cropping and geographic origin affect the genetic structure of *Plasmodiophora brassicae* populations. *Phytopathology* 104, 532-538.

Strelkov, S.E., Hwang, S-F., Manolii, V.P. et al. 2016. Emergence of new virulence phenotypes of *Plasmodiophora brassicae* on canola (*Brassica napus*) in Alberta, Canada. *Eur. J. Plant Pathol.* 145, 517-529.

Strelkov, S.E., Hwang, S-F., Manolii, V.P. et al. 2018. Virulence and pathotype classification of *Plasmodiophora brassicae* populations collected from clubroot resistant canola (*Brassica napus)* in Canada. *Can. J. Plant Pathol.* 40, 284-298.

Talbert, P.B. and Henikoff, S. 2020. What makes a centromere?. *Exp. Cell Res.* 389, 111895.

Tamada, T. and Asher, M.J.C. 2016. The Plasmodiophorid protist *Polymyxa betae*. In: Rhizomania (Biancardi, E. and Tamada, T. eds). Springer, Cham.

Tamada, T. and Kondo, H. 2013. Biological and genetic diversity of plasmodiophorid-transmitted viruses and their vectors. *J. Gen. Plant Pathol.* 79, 307-320.

Tanaka, S., Ito, S. and Kameya-Iwaki, M. 2001. Electron microscopy of primary zoosporogenesis in *Plasmodiophora brassicae*. *Mycoscience* 42, 389-94.

Thordal-Christensen, H. 2020. A holistic view on plant effector-triggered immunity presented as an iceberg model. *Cell. Mol. Life Sci.* 77, 3963-3976.

Tian, X., Wang, D., Mao, Z. et al. 2019. Infection of *Plasmodiophora brassicae* changes the fungal endophyte community of tumourous stem mustard roots as revealed by high-throughput sequencing and culture-dependent methods. *PLOS ONE* 14, e0214975.

Tomlinson, J.A. 1958. Crook root of watercress: III. The causal organism *Spongospora subterranea* (Wallr.) Lagerh. f.sp. *Nasturtii* f.sp.nov. *Trans Brit. Mycol. Soc.* 41, 491-498.

Tomlinson, J.A. and Hunt, J. 1987. Studies on watercress chlorotic leaf spot virus and on the control of the fungus vector (*Spongospora subterranea*) with zinc. *Ann. Appl. Biol.* 110, 75-88.

Tommerup, I.C. and Ingram, D.S. 1971. The life-cycle of *Plasmodiophora brassicae* Woron. in *Brassica* tissue cultures and in intact roots. *New Phytol.* 70, 327-332.

Toruño, T.Y., Stergiopoulo, I. and Coaker, G. 2016. Plant-pathogen effectors: cellular probes interfering with plant defenses in spatial and temporal manners. *Annu. Rev. Phytopathol.* 54, 419-441.

Tso, H.H., Galindo-González, L. and Strelkov, S.E. 2021. Current and future pathotyping platforms for *Plasmodiophora brassicae* in Canada. *Plants* 10, **1446.**

Upson, J.L., Zess, E.K., Białas, A. et al. 2018. The coming of age of EvoMPMI: evolutionary molecular plant-microbe interactions across multiple timescales. *Curr. Opin. Plant Biol.* 44, 108-116.

Vembar, S.S., Seetin, M., Lambert, C. et al. 2016. Complete telomere-to-telomere *de novo* assembly of the *Plasmodium falciparum* genome through long-read (>11 kb), single molecule, real-time sequencing. *DNA Res*. 23, 339-351.

Verma, G. and Surolia, N. 2018. Centromere and its associated proteins-what we know about them in *Plasmodium falciparum*. *IUBMB Life* 70, 732-742.

Walker, A.K. and Campbell, J. 2009. First records of the seagrass parasite *Plasmodiophora diplantherae* from the northcentral Gulf of Mexico. *Gulf Caribb. Res*. 21, 63-65.

Wallenhammar, A-C., Almquist, C., Schwelm, A. et al. 2014. Clubroot, a persistent threat to Swedish oilseed rape production. *Can. J. Plant Pathol.* 36, 135-141.

Wallenhammar, A.C. 1996. Prevalence of *Plasmodiophora brassicae* in a spring oilseed rape growing areas in central Sweden and factors influencing soil infestation levels. *Plant Pathol*. 45, 710-719.

Wang, Y., Tyler, B.M. and Wang, Y. 2019. Defense and counter defense during plant-pathogenic oomycete infection. *Ann. Rev. Microbiol*. 73, 667-696.

Williams, P.H. 1966. A system for the determination of races of *Plasmodiophora brassicae* that infect cabbage and rutabaga. *Phytopathol.* 56, 624-626.

Williams, P.H., Aist J.R. and Bhattacharya, P.K. 1973. Host-parasite relations in cabbage clubroot. In: Fungal pathogenicity and the plant's response (Byrde, R. J.W. and Cutting, C.V. eds). Academic Press, New York pp. 141-158.

Yadav, V., Yang, F., Reza, M.H. et al. 2019. Cellular dynamics and genomic identity of centromeres in cereal blast fungus. *mBio* 10, e01581-19.

Yu, F., Wang, S., Zhang, W. et al. 2019. Genome-wide identification of genes encoding putative secreted E3 ubiquitin ligases and functional characterization of PbRING1 in the biotrophic protist *Plasmodiophora brassicae*. *Curr. Genet.* 65, 1355-1365.

Zhai, Y., Mallik, I., Hamid, A. et al. 2020. Genetic diversity in potato mot-top virus populations in the United States and a global analysis of the PMTV genome. *Eur. J. Plant Pathol.* 156, 333-342.

Zhao, Y., Bi, K., Gao, Z. et al. 2017a. Transcriptome analysis of *Arabidopsis thaliana* in Response to *Plasmodiophora brassicae* during Early Infection. *Front. Microbiol.* 8, 673.

Zhao, Y., Gao, Z., Tian, B. et al. 2017b. Endosphere microbiome comparison between symptomatic and asymptomatic roots of *Brassica napus* infected with *Plasmodiophora brassicae*. PLOS ONE 12, e0185907.

Popular science summary

Clubroot is one of the most widely known diseases of plants in the mustard family, or *Brassicaceae* family, such as oilseed rape and cabbages. Enlarged roots or clubs are characteristic below-ground symptoms, often described as a "cabbage hernia" in many languages. Clubroot has a major impact on global production of *Brassica* oil crops and vegetables. It causes massive economical losses by directly reducing yields and quality and by having a long-term impact on infested soils. The causing agent of clubroot is *Plasmodiophora brassicae,* a soil-borne pathogen whose resting spores can survive up to two decades in soil. They are easily transmitted between fields and extremely resilient. There is no available commercial chemical control and *P. brassicae* has the ability to overpower resistant cultivars. Therefore, the clubroot pathogen represents a complex challenge for *Brassica* farmers all over the world.

P. brassicae is an intracellular pathogen that requires a living plant host. To provide necessary nutrients and a habitat for its own propagation, the pathogen applies various strategies of host manipulation. Many details of these strategies, including their cellular and molecular mechanisms are unknown. One of the main hinders for genome studies are difficulties in obtaining *P. brassicae* DNA, free from host and microbial contamination.

The general aim of this thesis was to enhance the genomic knowledge on *P. brassicae* and its intimate relationship with a plant host. We developed experimental procedures to extract large amounts of high-quality DNA from resting spores of the clubroot pathogen. This was a prerequisite for sequencing of *P. brassicae* genomes, which was done by using a so-called long-read technology. Sequencing is a process of determining the order of the four nucleotides Adenine, Thymine, Guanine and Cytosine in the DNA molecule. For example, a DNA sequence can be written as AATTTGGGCCCTA string of nucleotides. Often, the DNA of an entire genome is too long to be sequenced as one string. Instead, DNA is first broken into pieces which are sequenced and then assembled into the entire genome by using bioinformatics methods. This entire process can be compared with the assembly of a puzzle. It is less challenging and more accurate to assemble a smaller number of longer pieces than a larger number of shorter pieces. This is particularly true for repetitive sequences (e.g. ATTTTTTTAAAAA) that may seem to fit in several places within a puzzle. In other words, it is difficult to find their correct location.

Long-read sequencing of *P. brassicae* resolved highly repetitive genomic regions and allowed us, for the first time, to assemble contiguous sequences (contigs) into chromosomes. We generated 20 contigs, representing a nuclear genome with a size of 25.2 Mb. In addition, a mitochondrial genome (114.6 kb) was represented by a single sequence whose overlapping ends confirmed a circular structure of this genome. In total, long-read sequencing yielded 11.5% of repetitive sequences, indicating that the *P. brassicae* genome is more repeat-rich than previously estimated. We further characterized repetitive sequences and found that transposable elements (TEs) together with unclassified repeats were the most abundant group. They were distributed along the chromosomes, particularly clustered in telomeric and sub-telomeric regions and in large, repeat-rich islands. Whether these islands play a role of chromosomal centromeres remains to be further explored. We predicted 9,231 protein-coding genes in the nuclear genome. Roughly a half of these proteins lack information about their functional roles, mainly due low sequence similarity with proteins from other organisms. We were particularly interested to search for so-called effectors or small, secreted proteins which are involved in pathogen strategies to manipulate plant host defenses and metabolism. We identified 314 effector candidates and detected their enrichment with a motif that represents a shorter amino acid sequence, rich in valine, leucine and alanine. The effector candidates were distributed on all chromosomes, without significant clustering in repeat-rich sequences as seen in some other plant pathogens. However, more than a third of the predicted effectors were found in vicinity of TEs and unclassified repeats. This may be of importance since TEs are suggested to have roles in effector sequence diversification and regulation of gene expression.

Analysis of the *P. brassicae* mitochondrial genome revealed several unexpected findings. A previously unresolved 12 kb repetitive region and a gene-dense structure with presence of a larger number of introns that employ an atypical splicing mechanism were some of the peculiarities.

We believe that our findings together with *P. brassicae* whole genome sequences, available via a public database, will serve as a resource for further analysis and benefit the clubroot research community.

Populärvetenskaplig sammanfattning

Klumprotsjuka drabbar raps, kål och andra korsblommiga grödor tillhörande familjen *Brassicaceae*. Ett karaktäristiskt symptom är förstorade rötter. Sjukdomen har omfattande negativa konsekvenser för produktionen av oljeväxter och grönsaker världen över. Dels blir skördarna mindre och av lägre kvalitet, dels innebär infekterade jordar allvarliga långsiktiga effekter. Klumprotsjuka orsakas av patogenen *Plasmodiophora brassicae* vars sporer kan överleva i jord i upp till två decennier. Sporerna överförs dessutom lätt från plats till plats, och är extremt robusta. Bekämpningsmedel saknas, och det har visat sig svårt att skapa grödor som är resistenta mot *P. brassicae* över tid. Klumprotsjuka utgör således ett stort problem för växtodlande bönder världen över.

P. brassicae är en intracellulär patogen som fordrar en levande växtvärd. Patogenen har flera olika strategier att förmå värden att tillhandahålla näring och lämplig miljö. Detaljer kring dessa strategier är i många avseenden okända. En av de största utmaningarna för att kunna genomföra genomikstudier är att det är svårt att utvinna DNA från *P. brassicae* som inte är kontaminerat med värdens eller olika mikrobers DNA.

Syftet med denna avhandling var att via förbättrad kunskap om genomen i *P. brassicae* försöka klargöra hur samspelet mellan patogen och värd ser ut. Vi har därför utvecklat metoder att extrahera stora mängder DNA av hög kvalitet från sporer av klumprotpatogen. Detta var en förutsättning för att kunna sekvensera genom av *P. brassicae* med sekvenseringstekniker som resulterar i långa sekvenslängder. Med sekvensering avses en process där man fastställer i vilken ordning de fyra nukleotiderna Adenin, Tymin, Guanin och Cytosin kommer i DNA-molekylen. En kort DNA-sekvens kan exempelvis vara AATTTGGGCCCTA. Hela genomets DNA-sekvens är oftast för lång för att den ska vara möjlig att sekvensera i ett stycke. I stället måste den fullständiga sekvensen först delas upp i flera mindre bitar som sekvenseras var och en för sig, för att sedan fogas samman igen med hjälp av olika så kallade bioinformatiska metoder. Sammanfogandet kan liknas vid att lägga pussel; det är enklare att sätta ihop få och långa sekvenser än många korta, på samma sätt som det är enklare att lägga ett pussel med få stora bitar än ett med många små. Svårigheten gäller i synnerhet för repetitiva sekvenser (tex ATTTTTTTAAAAA) som kan passa på många olika ställen i pusslet vilket gör det extra svårt att finna deras rätta position.

Genom att använda sekvenseringstekniker som resulterade i långa sekvenslängder kunde 25,2 Mb kärngenom sammanfogas, liksom ett relativt stort genom på 114,6 kb representerande mitokondrien. Det senare genomet kunde sekvenseras i ett stycke och visade sig ha ändar som överlappade varandra vilket bekräftade mitokondriegenomets cirkulära struktur. Det sammanfogade kärngenomet bestod av 13 kromosomer avgränsade med telomerer i var sin ände, samt 7 kromosombitar med noll eller en telomersekvens. Av allt sekvenserat genom bestod 11,5% av repetitiva sekvenser vilket är en hög andel jämfört med vad som tidigare uppskattats. Vanligast var transposoner och oklassificerade repetitiva sekvenser. De återfanns framför allt i telomerregionerna samt i regioner med stora strukturella avvikelser. Vilken eventuell betydelse dessa regioner har för kromosomernas centromerer återstår att studera.

Vi kunde prediktera 9 231 gener i det nukleära genomet. För ungefär hälften saknades information om deras funktion, framför allt beroende på för låg överensstämmelse med motsvarande sekvenser hos andra organismer. Vi var särskilt angelägna om att använda genominformationen för att prediktera så kallade effektorkandidater, det vill säga gener som underlättar för patogenen att angripa och utvecklas i värdväxten. Totalt 314 proteiner i den kategorin identifierades varav en delmängd anrikade med sekvenser bestående av aminosyrorna valin, leucin och alanin. Effektorkandidater återfanns i alla kromosomer, och till skillnad från vissa andra växtpatogeners effektorer så kunde vi inte se några kluster i regioner med hög andel repetitiva sekvenser. Dock var mer än en tredjedel av kandidaterna placerade i närheten av transposoner eller oklassificerade repetitiva sekvenser. Detta kan vara viktigt eftersom transposonerna föreslagits ha betydelse bland annat för regleringen av effektorernas genuttryck.

Analyser av mitokondriegenom hos *P. brassicae* ledde till flera oväntade fynd, bland annat en repetitiv region om 12 kb som inte tidigare observerats i mitokondriegenomet och intronrika gener som använder en alternativ splitsningsmekanism.

Den nya genominformationen framtagen i detta projekt har skapat en värdefull grund för fortsatta analyser av, och ökad förståelse för, hur *P. brassicae* fungerar i samspelet med värdväxten. Inte minst torde patogenens sekvenserade genom – som finns tillgängligt i en öppen databas – kunna vara en viktig grund för framtida studier.

Acknowledgements

When I grew up, borders between people and countries became increasingly important. Education was one of the rare and not always easily accessible pathways to overcome these obstacles. My journey towards PhD education made it possible for me to cross many visible and hidden borders. Most importantly, I met kind and generous people without whose friendship, support and care I would not be who I am today. To every one of you - I am deeply grateful!

I would like to express my special thanks to:

My main supervisor Christina Dixelius for accepting me as a PhD student, even though I was an outsider with the background in animal science. You opened a door to more than 140 years long, well-founded fascination with Plasmo and I really enjoyed exploring, for me, a whole new world. Thank you for sharing your extensive knowledge and "outside of the box" discussions. I have learned a lot because you allowed me to find my own way and because you challenged my views. Your quick responses, guidance with thesis writing, support and "nothing is impossible" attitude are very appreciated!

Johan Fogelqvist, for co-supervising me during my first year and for introducing me to Plasmo data. Your analyses and the time you took to answer my e-mails and provide input during all these years were essential for the work presented in this thesis!

Åke Olsson, thank you for accepting to be my co-supervisor and for valuable and constructive feedback.

Members of my PhD evaluation committee: Anders Hafrén, Anki Rönnberg-Wästljung, Hanna Johannesson and Pär Ingvarsson. Thank you for your comments, advice and support.

Olga Vinnere Pettersson and Christian Tellgren-Roth, your support with sequencing and assembly data is highly appreciated! Jacques Daiant, thank you for extensive annotation help and for patiently answering many e-mails. Johan Nylander, your input about phylogenetic analyses was very valuable.

All previous and present colleagues at the Plant Biology Department: thank you for always being helpful and for all the smiles in the corridors and chats at the coffee corners.

A special thanks to "Dixan group" former (Anna T., Anna Å., Georgios, Hanneke, Kristian, Louise, Ravi, Sara, Sultana, and Zhen) and present members (Fredrik, Shailja, Shridhar, Tua and Xinyi). I am very grateful for your friendship, teamwork, scientific and non-scientific exciting discussions through all these years. Anna Å. and Georgios, you made me feel welcome from my first day in the group. Your help and guidance have been vital for me, and your care and friendship are truly appreciated. Kristian, thank you for sharing scripts and protocols, candy supplies and laughs! Most of all, thank you for your kindness and support throughout the years! Sara, our protocol would not have become a publication without you. Thank you for all your dedicated and meticulous work! Fredrik, thank you for all your help and good luck with your future Plasmo projects!

Khushwant, you taught me how to extract my first Plasmo spores and manage not so appealing smell of mushy clubroots. Thank you for being an excellent teacher and for sharing your extensive wet lab skills!

Lotta, Monica and Qing, thank you for taking care about all administrative work and Anna S. for well-timed deliveries. Per and Urban, your positive attitude and great care about plant materials meant a lot. Björn and Christian, your excellent support with computer-related issues allowed me to work remotely and complete my analysis and thesis writing.

Thanks to the effort of a larger number of colleagues and students, I was able to experience a rewarding teaching and supervising role. A special thanks to Minerva for being a great example of a dedicated teacher, I truly enjoyed our time with students and learned a lot. Anki, thank you for all your support as a course leader! Elen, thank you for your devoted work on a master project that generated hundreds of clubroots! Peter, your help with Arabidopsis mutants is highly appreciated!

Kanita, one of the nicest things at the beginning of my PhD studies was to get to know you as a colleague and a friend. I cannot thank you enough for your care and support!

Dušica and Marijana, your long-lasting friendships are among the most cherished treasures I have. Linda, I am privileged to have you as a friend.

My love to my Serbian family. I miss you so much!
My Swedish family - thank you for caring about me!

And finally, to Per for your love, care, support, patience and knowledge.
For believing that I can and for all your help.
For being exactly who you are!

Root Gall Formation, Resting Spore Isolation and High Molecular Weight DNA Extraction of *Plasmodiophora brassicae*

Sara Mehrabi#, *, Suzana Stjelja# and Christina Dixelius

Swedish University of Agricultural Sciences, Department of Plant Biology, Uppsala BioCenter, Linnean Center for Plant Biology, Uppsala, Sweden

*For correspondence: sara.mehrabi@slu.se

#Contributed equally to this work

[Abstract] Isolation of DNA from obligate biotrophic soil-borne plant pathogens is challenging. This is because of their strict requirement of living plant tissue for their growth and propagation. A soil habitat further imposes risk of contamination from other microorganisms living in close vicinity of the plant roots. Here we present a protocol on how to prepare DNA suitable for advanced molecular analysis on the soil-borne pathogen *Plasmodiophora brassicae,* a peculiar unicellular plant pathogenic organism, causing disease on Crucifers. First, it is important to grow *Brassica* or *Arabidopsis* plants in infested soils below a temperature of 25 °C under moist conditions to promote root gall formation. Root galls should be harvested ahead of initiation of the decomposing process, no later than four or nine weeks post inoculation of *Arabidopsis* or *Brassica* plants, respectively. Resting spores with reduced numbers of soil organisms are achieved by gradient centrifugations of homogenized gall tissues. Treatments with 70% alcohol and a suit of different antibiotics promote *P. brassicae* purity. A CTAB-based procedure allows isolation of high quality DNA suitable for massive parallel sequencing analysis.

Keywords: *Arabidopsis*, *Brassica*, Clubroot, DNA, *Plasmodiophora brassicae*, Resting spores, Rhizaria

[Background] *Plasmodiophora brassicae* is a soil-borne plant pathogen causing root galls (clubs) in the *Brassicaceae* family including *Arabidopsis*. The clubroot disease has a major impact on oilseed rape (canola) and cabbage cultivation worldwide. *P. brassicae* is an obligate biotroph (require a host for growth) assigned to the supergroup Rhizaria, one of the least studied organism groups of eukaryotes (Sierra *et al*., 2016; Sibbald and Archibald, 2017). Phylogenetically, *P. brassicae* belongs to a plant pathogenic group of protists in Phytomyxea (Neuhauser *et al*., 2011 and 2014; Adl *et al*., 2012). Few genomes of related species are available, a circumstance which has considerably delayed the molecular analysis and genome comparisons. *P. brassicae* forms hardy resting spores in the clubs, spores that have the capacity to remain dormant for decades in the soil, ready for new rounds of root infections if a host plant grow nearby. Here we describe how to generate diseased plants, isolate resting spores from root galls followed by extraction of large amounts of DNA. This protocol is a further improvement and clarification of the procedures described in Schwelm *et al*. (2015). The outlined work is substantial but yields high-quality DNA suitable for long-read massive parallel sequencing.

Materials and Reagents

A. Materials

1. Safety glasses, gloves and lab coat
2. Filter paper
3. Plant pots, small pots (6 x 6 x 5 cm) and big pots (13 x 13 x 13 cm)
4. Plant trays (34 x 22 x 4 cm)
5. Soil (S-soil, Hasselfors Garden, Örebro, pH 5.5-6.5) composed of sighted light peat, black peat, perlite, sand, and lime
6. Petri dishes 10 cm (ø)
7. Miracloth Calbiochem® (Merck, catalog number: 475855)
8. Tubes (Falcon tube 15 ml, SARSTEDT, catalog number: 62.554.001; Falcon tube 50 ml, SARSTEDT, catalog number: 62.547.004; Eppendorf micro-tube 2 ml, SARSTEDT, catalog number: 72.695.500)
9. Plastic and glass beakers
10. Scalpel
11. Mortar, pestle and spoon (sterile and pre-chilled)
12. Filtropur S0.2 (SARSTEDT, catalog number: 83.1826.001)

B. Plants

1. *Brassica rapa* cv. 'Granaat' (European Clubroot Differential Set ECD-05)
2. *Arabidopsis thaliana* Col-0

C. Plasmodiophorid

Plasmodiophora brassicae strain e3 (Fähling *et al*., 2004; the strain is available upon request to the authors)

Note: Not all strains incite disease on Arabidopsis.

D. Molecular biology working kit

Spectrum^{TM} Plant Total RNA Kit (Sigma-Aldrich, catalog number: STRN50)

E. Other reagents

*Note: *Those solutions are prepared with sterile distilled water.*

Spore isolation

1. Ethanol (70%, 500 ml*)
2. Sodium hypochlorite (1%, 500 ml*) (Commercial bleach)
3. Sterile distilled water 5 x 1L
4. Ficoll PM 400 (16% and 32%*) (Sigma-Aldrich, catalog number: F4375)

5. Rifampicin (100 mg/ml stock, prepared with methanol solvent) (Duchefa Biochemie, catalog number: R0146)
6. Streptomycin sulfate (100 mg/ml stock*) (Thermo Fisher Scientific, catalog number: 11860038)
7. Carbenicillin disodium (100 mg/ml stock, prepared with water solvent) (Duchefa Biochemie, catalog number: C0109)
8. Pimaricin (100 mg/ml stock*) (Merck, Sigma-Aldrich, catalog number: 1.07360)
9. Cefotaxime sodium (100 mg/ml stock*) (Duchefa Biochemie, catalog number: C0111)
10. Hygromycin B (50 mg/ml stock) (Duchefa Biochemie, catalog number: H0192)
11. Lysozyme from chicken egg white (4 mg/ml*) (Sigma-Aldrich, catalog number: L6876)
12. Sodium chloride (NaCl) (Sigma-Aldrich, catalog number: 31434)
13. Potassium chloride (KCl) (Merck, catalog number: 104936)
14. di-Sodium hydrogen phosphate dihydrate (Na_2HPO_4) (Merck, catalog number: 119753)
15. Potassium dihydrogen phosphate (KH_2PO_4) (Merck, catalog number: 104873)
16. Tris-HCl (Trizma® hydrochloride solution) (1 M, pH 7.5) (Sigma-Aldrich, catalog number: T2694)
17. DNase I, RNase-free (Thermo Fisher Scientific, Thermo Scientific™, catalog number: EN0521)
18. Proteinase K (20 mg/ml stock*) (Sigma-Aldrich, catalog number: RPROTK-RO)
19. EDTA (0.5 M) (VWR, catalog number: 20294.294)
20. N-lauroylsarcosine sodium salt solution (1%, v/v) (Sigma-Aldrich, catalog number: 61747)
21. 1x PBS buffer (see Recipes)
22. 1x TE buffer (pH 7.5) (see Recipes)
23. Termination buffer (see Recipes)

DNA extraction

1. Liquid nitrogen
2. Tris-HCl (Trizma® hydrochloride solution) (1 M, pH 7.5) (Sigma-Aldrich, catalog number: T2694)
3. EDTA (VWR, catalog number: BDH9232)
4. Sodium chloride (NaCl) (5 M stock *) (Sigma-Aldrich, catalog number: 31434)
5. Hexadecyltrimethylammonium bromide (CTAB) (Sigma-Aldrich, catalog number: H6269)
6. 2-Mercaptoethanol (VWR, catalog number: 436022A)
7. Phenol:chloroform:isoamyl alcohol 25:24:1 (pH 7.5-8.0) (Carl Roth, catalog number: A156.2)
8. RNAse A, DNase and protease-free (10 mg/ml) (Thermo Fisher Scientific, Thermo Scientific™, catalog number: EN0531)
9. Chloroform, EMSURE® ACS, ISO, Reag. Ph. (Merck, catalog number: 1.02445.1000)
10. Ethanol (95%)
    Note: Pre-chill at -20 °C before use.
11. Ethanol (70%)*
12. 0.1x TE buffer (see Recipes)

13. CTAB extraction buffer (see Recipes)

F. Media and buffers (see Recipes)
1. 1x PBS buffer
2. 1x TE buffer
3. Termination buffer
4. CTAB extraction buffer

Equipment

1. Growth chamber (Percival AR82L2/Split) or greenhouse
2. Analytic balance (Mettler Toledo, model: AE100)
3. Household mixer (Rusta, catalog number: 90951442, Max power 170 W)
4. Liquid nitrogen container
5. Water bath (JULABO, model: Julabo TW12, catalog number: 9550112)
6. pH meter (Mettler Toledo, model: SevenCompact S220, catalog number: 30019028)
7. Tabletop centrifuge for Eppendorf tubes (Thermo Fisher Scientific, Thermo Scientific^{TM}, model: Heraeus^{TM} Fresco^{TM} 17, catalog number: 75002420)
8. Tabletop centrifuge for Falcon-tubes (Eppendorf, model: 5804/5804 R, catalog number: 5805000327)
9. NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA)
10. Microscope (Zeiss, Axioplan; camera: Leica Microsystems, model: Leica DFC295)
11. Stereoscope (Lecia, model: MZ FL III)
12. Autoclave

Procedure

A. Preparation of infested soil
1. Start preparing at least two rounds of clubroots of chosen strain or pathotype for soil infestation well in time before any larger experiments. This is achieved by crushing soft clubroots followed by mixing the tissue thoroughly with wet soil.
2. Grow *Brassica* plants for 7 to 9 weeks or *Arabidopsis* plants for 3 to 4 weeks in the infested soil. Repeat this step. Clubs can also be left in the soil to decompose but a mixing step speeds up the procedure. Spores per gram soil should exceed 100,000 which can be determined by qPCR analysis (Wallenhammar *et al*., 2012). Store infested soil at -20 °C for later use.

B. Preparation of diseased *Brassica* or *Arabidopsis* plants
1. Place *Brassica* seeds on a wet filter paper in a Petri dish at room temperature, 8-10/14-16 h light/dark regime to germinate (1-2 days).

2. Transfer 2-3 germinated seeds to a large pot with infested soil. Cultivate the plants for eight to nine weeks under greenhouse conditions (16 h light/8 h dark cycle and 22 °C day/18 °C night temperature). Add nutrients (2 ml Blomstra, Cederroth, Upplands Väsby/l water) on a daily basis.
3. Collect clubs when just started to get soft (Figures 1A to 1D), rinse them carefully with water and store at -20 °C for later use.

   Note: Clubs younger than 7-9 weeks old are too firm for homogenization by a mixer, limiting spore isolation, while older clubs left in soil quickly start to decompose yielding few spores. Weekly monitoring of plants and club development by uprooting is important because club formation and 'rate of maturity' can vary between experiments.

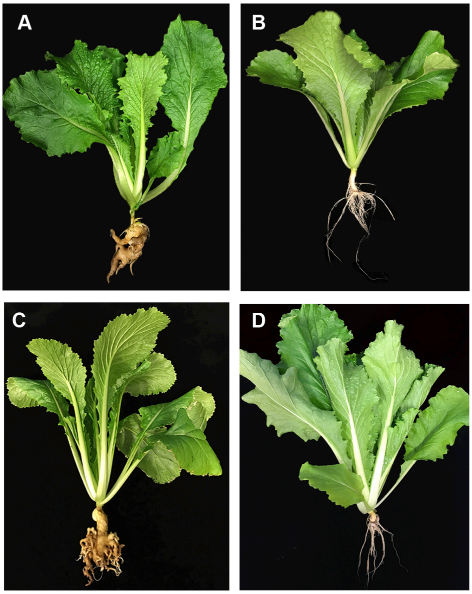


Figure 1. *Brassica rapa* cv. Granaat. A and B. Clubs harvested after 5 weeks in *P. brassicae* infested soil (A) and the corresponding control plant (B). C and D. Clubs harvested after 7 weeks in *P. brassicae* infested soil (C) and the corresponding control plant (D).

4. Alternatively, incubate *Arabidopsis* Col-0 seeds in sterile water overnight at 4 °C.
5. Plant five seeds per small pot of un-infested soil covered by transparent plastic top and keep them in short day conditions (16 h dark/8 h light cycle at 22 °C and 60% relative humidity (RH) for 14 days).
6. Transfer 24 plants into a tray with infested soil (Figure 2A) and cover the tray with a transparent plastic top. Keep the trays in short day conditions, at 15 °C, and 60% RH. Remove the lid after 4 days.

7. Galls are visible after 3 to 4 weeks (Figures 2B to 2H). Harvest and rinse the clubs carefully, and store at -20 °C for later use.

   Note: Clubs can be obtained from most crops, weedy and wild species in the Brassicaceae family.

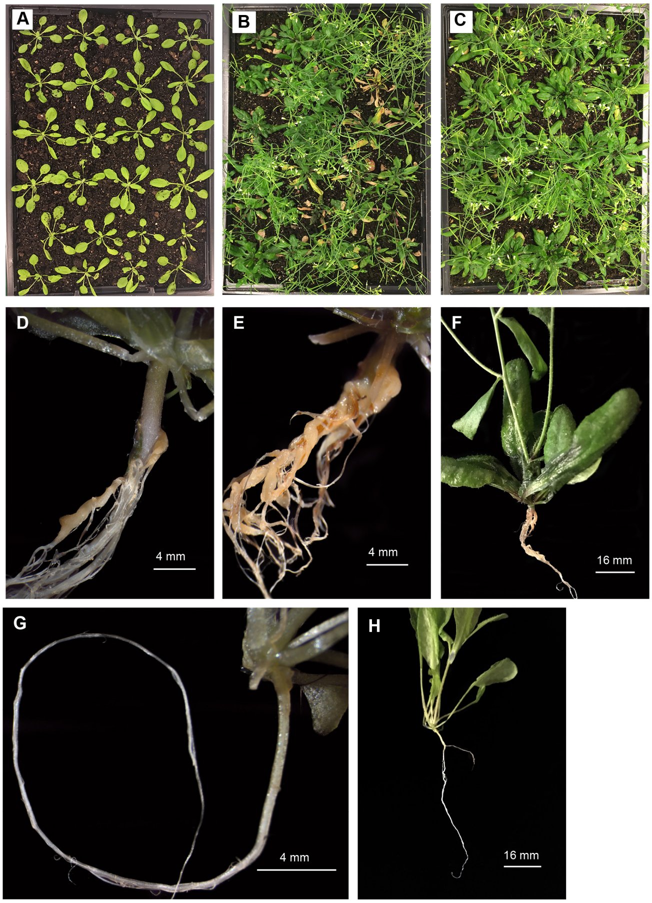


Figure 2. *Arabidopsis thaliana* Col-0. A. Healthy plants after two weeks; B. Diseased plants after four weeks in *P. brassicae* infested soil; C. Control plants, grown four weeks in H_2O treated soil; D-F. Gall formation on *Arabidopsis* roots; G and H. Control, *Arabidopsis* root grown in H_2O treated soil.

C. Isolation of resting spores, Day 1 (Steps C1 to C13) and Day 2 (Steps C14 to C20)

1. Use 250-300 g frozen clubs. Thaw shortly at room temperature. Surface sterilize the material in 70% ethanol for 2 min followed by 1% sodium hypochlorite for 5 min. Rinse 5 times with sterile H_2O. Repeat the entire surface sterilization step.

Note: Only use sterilized H_2O from now and onwards. Autoclave Miracloth and rinse the household mixer with 70% alcohol and sterile H_2O before use.

2. Homogenize galls in 500 ml H_2O using a household mixer. Avoid over-heating of the material by running the mixer for 30 sec followed by a 30 sec break. Repeat until a colloidal suspension has formed (Figure 3A).

   Note: Scalpel can be used to cut the root material into smaller pieces ahead of the mixing.
3. Filter the homogenized tissue through 4 layers of Miracloth to remove debris. Depending on the club size and amount of debris this step may require a new round of filtering (Figure 3A).
4. Transfer the filtered liquid into 50 ml Falcon tubes up to maximum volume and spin.

   Note: All centrifugations should be carried out using a tabletop centrifuge at 3,650 x g for 15 min if not otherwise is mentioned.
5. A pellet with two layers will form: a brown upper layer containing spores and a white layer with starch (Figure 3B). Carefully remove the supernatant with a pipet using a 1 ml tip. Transfer the spores to a new tube. Avoid starch contamination.

   Note: A dense spore layer may require a spoon to facilitate spore removal.
6. Wash the spores by re-suspending them in about 40 ml H_2O and repeat centrifugation (Figure 3C). Repeat washing and centrifugation if a white layer of starch is visible.

   Note: All volumes from now and onwards are adjusted to 2-4 ml of spores.

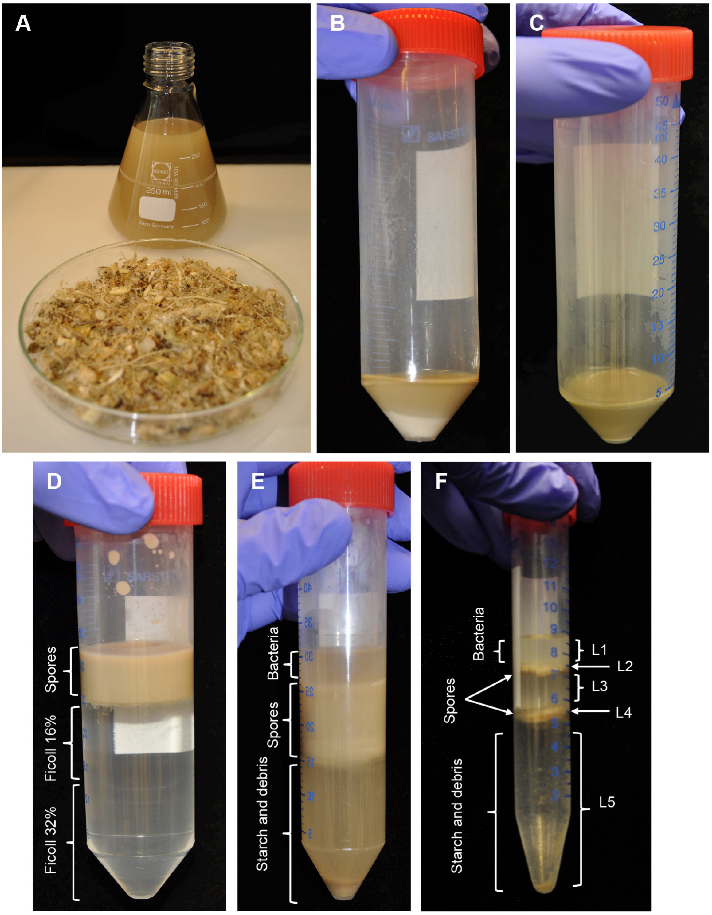


Figure 3. Spore isolation. A. Filtered liquid containing spores and the club roots left over after Miracloth filtering. B. A pellet with two layers: spores (brown) and starch (white). C. Spore layer after washing steps. D. Spores from Step C7 on top of two-step Ficoll gradient. E. Three layers after Ficoll gradient centrifugation (Step C9): upper layer with bacteria, middle layer mainly spores and bottom layers with starch and debris. F. Five layers (L1 to L5) after Ficoll gradient (Step C16), upper layer (L1) with bacteria, middle layers mainly spores (L2 and L4) and bottom layer (L5) with starch and debris.

7. Re-suspend the pellet in 5 ml H_2O. Check the purity and presence of the spores using a microscope (Figure 4A).

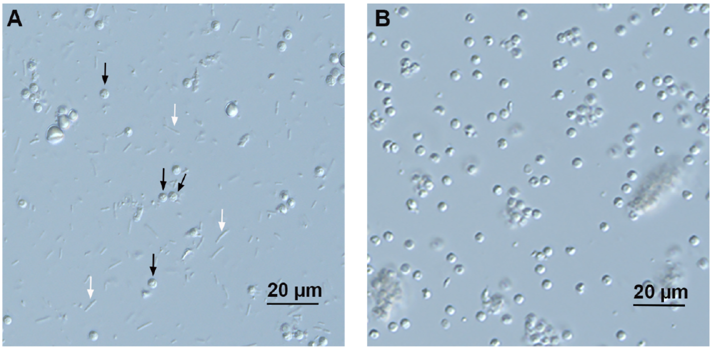


Figure 4. Different levels of spore purity. A. Spore preparation at Step C7. Spores are indicated with black arrows and debris is indicated with white arrows. B. Clean spores got at Step C20. Spore diameter, 2.75-3.75 µm.

8. Prepare a density gradient in a 50 ml Falcon tube. Add 32% Ficoll over which carefully layer 16% Ficoll and add the re-suspended spores from Step C7 on top (3v 32% Ficoll:2v 16% Ficoll:1v spores). Make sure the interphase between 3 layers is undisturbed (Figure 3D).
9. Centrifuge for 15 min at 400 *x g*. Avoid high-speed centrifugation. Several layers will form (Figure 3E).
10. Carefully aspirate and discard the upper layer. Transfer the spores to a 15 ml Falcon tube. Wash the spores using 10 ml of H_2O and centrifuge.
11. Dissolve the pellet in H_2O reaching a final volume of 5 ml. Add 5 µl of rifampicin and streptomycin each into the tube and incubate for 1 h at 37 °C. Pellet the spores and wash them with 10 ml of H_2O. Repeat centrifugation and remove supernatant.
12. Treat the spores with 5 ml ethanol (70%) for 2 min (not longer) and centrifuge. Wash the spores with 10 ml of H_2O and centrifuge. Repeat the washing-step with H_2O at least twice.
    Note: It is important to remove all ethanol.
13. Suspend the pellet in 5 ml H_2O. Add 5 µl carbenicillin, 5 µl pimaricin, 12.5 µl cefotaxime and 2.5 µl hygromycin B. Incubate the tube at room temperature overnight.
    Note: The antibiotics are active against various bacteria and fungal species.
14. Pellet the spores, wash with 10 ml of H_2O and repeat centrifugation. Re-suspend the spores in 5 ml lysozyme solution and incubate for 2 h at 37 °C.
15. Pellet the spores and re-suspend the pellet in 5 ml PBS buffer.
16. Repeat Step C8 and Step C9 with Ficoll density gradient centrifugation.
17. Collect the spores, layers 2 and 4 (Figure 3F) and wash them with 10 ml H_2O and centrifuge. Dissolve the pellet in 2 ml TE buffer containing 20 µl of DNase I and incubate for 2 h at 37 °C.
18. Add 3 ml termination buffer and incubate for 4 h at 37 °C.
19. Pellet and wash the spores 3 times each using 10 ml H_2O.
20. Dissolve the pellet in half volume of H_2O. Store the spores at -20 °C for future DNA or RNA

extractions. Clean spores should now have been achieved (Figure 4B).

Note: The amount of water to dissolve the pellet is dependent on the amount of spores. If we end up with 500 µl spores, we add 250 µl water.

D. DNA isolation

1. Add 7 ml of CTAB extraction buffer and 21 µl of 2-mercaptoethanoethanol (0.3% v/v) to a 15 ml Falcon tube.
2. Pre-heat the mixture in a water bath at 65 °C.
3. Grind the frozen spores to a fine powder in liquid nitrogen. Use mortar and pestle.
4. Add the spore-powder into the pre-heated extraction buffer (about 300 mg of spore-powder in 7 ml extraction buffer).
5. Incubate the solution at 65 °C and mix gently every 10 min for 1 h.

   Note: Avoid mechanical disruption of DNA by vortexing and excessive pipetting! This is very important throughout the procedure.
6. Add 1 volume of phenol:chloroform:isoamyl alcohol and mix gently for 5 min. Centrifuge and pipette the upper aqueous phase into a new Falcon tube.

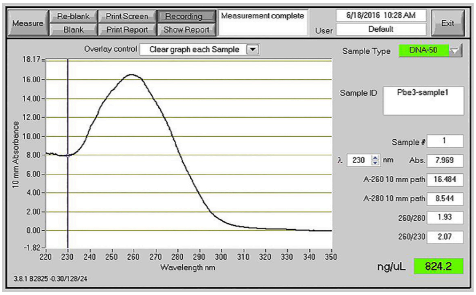
   Note: Avoid to pipet aqueous/organic layer interface.
7. Add 5 µl RNase A to the solution and incubate for 45 min at 37 °C. Mix gently by inverting tubes 4-6 times.
8. Add 1 volume of phenol:chloroform:isoamyl alcohol and mix by inverting for 5 min. Centrifuge and pipette the aqueous phase to a new Falcon tube.

   Note: All centrifugations should be carried out on a tabletop centrifuge at 3,650 x g for 10 min if not otherwise mentioned.
9. Add 1 volume of chloroform and mix by inverting for 5 min. Centrifuge and pipette the upper aqueous phase to a new Falcon tube.

   Note: Repeat this step if phenol still is present in the solution.
10. Add 1/2 volume of 5 M NaCl to the sample and mix gently by inverting. Add 3 volumes of fresh and cold ethanol (95%) and mix gently by inverting. Keep at -20 °C for 1 h, not longer.
11. Centrifuge to pellet DNA. Carefully decant the supernatant and wash (do not re-suspend) the DNA pellet with 3 ml of 70% ethanol. Centrifuge again and carefully decant the supernatant.

    Note: Repeat washing to remove precipitated NaCl and other contaminants were seen as whitish crystals in the pellet.
12. Remove all visible ethanol from the final pellet. Let air-dry the DNA for 15 min at room temperature. Carefully suspend DNA in 200 µl (or less) of 0.1x TE buffer. Optional: dissolve DNA overnight at +4 °C.

    Note: Handle DNA with care, avoid vortexing and heavy pipetting. If DNA will be used for massively parallel sequencing, adjust elution buffer according to planned procedures.
13. Assess the quality and concentration of extracted DNA by using a NanoDrop or DropSense or a similar device (Figure 5).



Measurement complete

6/18/2016 10:28 AM

User Default

Overlay control: Clear graph each Sample

Sample Type: DNA-50

Sample ID: Pbe3-sample1

Sample #: 1

λ 230 nm Abs. 7.969

A-260 10 mm path: 16.484

A-280 10 mm path: 8.544

260/280: 1.93

260/230: 2.07

ng/uL: 824.2

Figure 5. DNA quality check. NanoDrop measurement profile of *P. brassicae* DNA using this protocol.

E. RNA isolation

1. Grind the frozen spores to a fine powder in liquid nitrogen. Use a mortar and pestle.
2. Continue RNA extractions using the Spectrum$^{\text{TM}}$ Plant Total RNA Kit (Sigma-Aldrich) according to the manufacturer's instructions.

Notes

Root infection varies much due to the environmental conditions and experiments should first be run to ensure proper club formation. The yield of resting spores and consequently of *P. brassicae* DNA vary depending on the club size and stage of maturation.

Recipes

1. 1x PBS buffer (50 ml)
   137 mM NaCl
   2.7 mM KCl
   10 mM Na_2HPO_4
   1.8 mM KH_2PO_4
   Sterile water up to the final volume
   Adjust pH to 7.4 with HCl and autoclave
2. 1x TE buffer (30 ml)
   10 mM Tris-HCl
   0.1 mM EDTA
   Sterile water up to final volume, autoclave
   Store at 4 °C or room temperature
3. Termination buffer (50 ml)
   0.5 M EDTA
   1% N-lauroylsarcosine (v/v)

0.1 mg/ml Proteinase K

Note: The Proteinase K powder should be dissolved at a concentration of 20 mg/ml in sterile 50 mM Tris (pH 8.0), 1.5 mM calcium acetate and stored at -20 °C.

4. CTAB extraction buffer (CTAB) (50 ml)
   100 mM Tris-HCl (pH 7.5)
   25 mM EDTA
   1.5 M NaCl
   2 % (w/v) CTAB (cetyltrimethylammounium bromide)
   Sterile water up to final volume
   Filter sterilize using Filtropur and store at room temperature

Acknowledgments

We thank the plant growth facility staff at the Faculty of Natural Resources and Agricultural Sciences, Swedish University of Agricultural Sciences (SLU) for close attention of the plants. We also thank Dr. Arne Schwelm that started the extensive work of *P. brassicae* DNA preparation 2009. The authors declare that they have no conflicts of interest or competing interests. The work was funded by the Swedish Research Council (Formas) and SLU.

References

1. Adl, S. M., Simpson, A. G., Lane, C. E., Lukes, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Le Gall, L., Lynn, D. H., McManus, H., Mitchell, E. A., Mozley-Stanridge, S. E., Parfrey, L. W., Pawlowski, J., Rueckert, S., Shadwick, L., Schoch, C. L., Smirnov, A. and Spiegel, F. W. (2012). The revised classification of eukaryotes. *J Eukaryot Microbiol* 59(5): 429-493.
2. Fähling, M., Graf, H. and Siemens, J. (2004). Characterization of a single-spore isolate population of *Plasmodiophora brassicae* resulting from a single club. *J Phytopathol* 152(7): 438-444.
3. Neuhauser, S., Kirchmair M., Bulman, S. and Bass, D. (2014). Cross- kingdom host shifts of phytomyxid parasites. *BMC Evol Biol* 14: 33.
4. Neuhauser, S., Kirchmair, M. and Gleason, F. H. (2011). Ecological roles of the parasitic phytomyxids (plasmodiophorids) in marine ecosystems - a review. *Mar Freshw Res* 62(4): 365-371.
5. Schwelm, A., Fogelqvist, J., Knaust, A., Julke, S., Lilja, T., Bonilla-Rosso, G., Karlsson, M., Shevchenko, A., Dhandapani, V., Choi, S. R., Kim, H. G., Park, J. Y., Lim, Y. P., Ludwig-Muller, J. and Dixelius, C. (2015). The *Plasmodiophora brassicae* genome reveals insights in its life cycle and ancestry of chitin synthases. *Sci Rep* 5: 11153.

6. Sibbald, S. J. and Archibald, J. M. (2017). More protist genomes needed. *Nat Ecol Evol* 1(5): 145.
7. Sierra, R., Canas-Duarte, S. J., Burki, F., Schwelm, A., Fogelqvist, J., Dixelius, C., Gonzalez-Garcia, L. N., Gile, G. H., Slamovits, C. H., Klopp, C., Restrepo, S., Arzul, I. and Pawlowski, J. (2016). Evolutionary origins of rhizarian parasites. *Mol Biol Evol* 33(4): 980-983.
8. Wallenhammar, A. C., Almquist, C., Söderström, M. and Jonsson, A. (2012). In field distribution of *Plasmodiophora brassicae* measured using quantitative real-time PCR. *Plant Pathol* 61(1): 16-28.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

natureresearch

OPEN

The architecture of the *Plasmodiophora brassicae* nuclear and mitochondrial genomes

Suzana Stjelja[1], Johan Fogelqvist[1], Christian Tellgren-Roth[2] & Christina Dixelius[1*]

***Plasmodiophora brassicae* is a soil-borne pathogen that attacks roots of cruciferous plants causing clubroot disease. The pathogen belongs to the Plasmodiophorida order in Phytomyxea. Here we used long-read SMRT technology to clarify the *P. brassicae* e3 genomic constituents along with comparative and phylogenetic analyses. Twenty contigs representing the nuclear genome and one mitochondrial (mt) contig were generated, together comprising 25.1 Mbp. Thirteen of the 20 nuclear contigs represented chromosomes from telomere to telomere characterized by [TTTTAGGG] sequences. Seven active gene candidates encoding synaptonemal complex-associated and meiotic-related protein homologs were identified, a finding that argues for possible genetic recombination events. The circular mt genome is large (114,663 bp), gene dense and intron rich. It shares high synteny with the mt genome of *Spongospora subterranea*, except in a unique 12 kb region delimited by shifts in GC content and containing tandem minisatellite- and microsatellite repeats with partially palindromic sequences. *De novo* annotation identified 32 protein-coding genes, 28 structural RNA genes and 19 ORFs. ORFs predicted in the repeat-rich region showed similarities to diverse organisms suggesting possible evolutionary connections. The data generated here form a refined platform for the next step involving functional analysis, all to clarify the complex biology of *P. brassicae*.**

Unicellular eukaryotes or protists can be found in any habitat worldwide and fossil records are available for certain taxa where some are dated as early as the Precambrian period[1]. Among the protists, Rhizaria is a large and diverse organism group in the kingdom Chromista which has experienced a number of taxonomic re-evaluations[2]. A few plant pathogens can be found here among which *Plasmodiophora brassicae* is the most well known, located in the Phytomyxea class[3]. This plasmophorid pathogen attacks roots of numerous cruciferous plant species resulting in typical swollen roots (clubs or galls), giving rise to the disease name. The clubroot disease has been known about 100 years, and is now present in more than 60 countries worldwide[4–6]. *P. brassicae* is an obligate biotroph and has a strict requirement of host tissue for growth and multiplication. The resting spores of *P. brassicae* can survive in soil for many years, before hatching under suitable conditions followed by zoospore release. The host plant infection process is divided into two phases. A short stage where root hairs are infected by zoospores in the soil, followed by formation of primary plasmodia and second round of zoospores. Upon release, these zoospores, initiate a new round of infection where large intracellular plasmodia in root cortical cells develops[7,8]. Genetic recombination is thought to take place during zoospore development and early infection phases but these events and possible reproduction stages are not completely understood[9]. Further details on biology and other aspects on *P. brassicae* are covered elsewhere[10–12].

Very restricted genome information is available from organisms in Rhizaria. This is because they commonly colonize complex ecological niches from which pure DNA is difficult to retrieve in required quantities. Besides *P. brassicae*, nuclear data today derive from the foraminifera *Reticulomyxa filosa*, the chlorarachniophyte alga *Bigelowiella natans*, the plasmophorid *Spongospora subterranea* and transcriptome datasets on marine species[13–18]. *S. subterranea* is another soil-borne pathogen causing the potato powdery scab disease but this organism can also act as a vector for the Potato Mop Top Virus[19]. *P. brassicae* and *S. subterranea* are the two most closely related plasmodiophorids based on present sequence information[20].

[1]Department of Plant Biology, Uppsala BioCenter, Linnéan Center for Plant Biology, Swedish University of Agricultural Sciences, P.O. Box 7080, SE-75007, Uppsala, Sweden. [2]Uppsala Genome Center, Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, BMC, Box 815, SE-751 08, Uppsala, Sweden. *email: Christina.Dixelius@slu.se

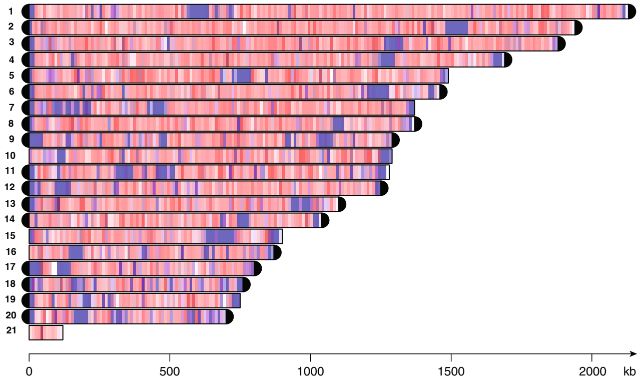


Figure 1. Contigs and telomeres in *Plasmodiophora brassicae* e3 strain. Contig sizes of the nuclear (no. 1 to 20) and mitochondrial genome (no. 21). Telomeres are marked with black ends. Density of coding (red) and repeat (blue) sequences in non-overlapping sliding 10 kbp windows. The color intensity is proportional to the given feature density.

Application of long-read sequencing greatly supports not only reliable pathotyping (diagnostics) or identification of important structural variants when genomic gaps with repetitive or unique sequences are closed but also accurately resolve chromosomes in complex genomes. Genomic information of high quality is crucial for enhancing our understanding of important evolutionary events, which could give us insights into related organisms together with events behind lost and gained traits. Such information could in case of plant pathogens be extra valuable for development of durable control strategies. Reliable genomic information is essential not least for experimentally challenging obligate biotrophic soil-borne pathogen as *P. brassicae* where a number of questions in its life-cycle and differences between pathotypes remain to be clarified.

We previously made a *de novo* genome assembly of the *P. brassicae* strain e3 based on a combination of Illumina and 454 sequencing approaches[16]. This achievement was followed by genome information for isolates from Canada, China and Germany based on similar technologies[21–23]. The former sequencing of the *P. brassicae* e3 genome[16] generated two mitochondrial contigs whose complete assembly could not be achieved, thus the mitochondrial sequence was not made public. Here, we applied long-read PacBio RSII single molecule real-time (SMRT) sequencing technology to fill in sequence gaps and further resolve regions with repetitive sequences of the *P. brassicae* e3 genome. We present new data based on twenty contigs representing the nuclear genomic content of *P. brassicae* e3 and one contig covering the entire mitochondrial genome generated by this approach. We were able to identify telomere sequences similar to those found in green algae and *Theileria annulata*, an apicomplexan parasite. The assembly and *de-novo* annotation of the mitochondrial genome revealed a 114,663 bp large genome with a complex sequence organization and a distinct 12 kb repeat-rich region. These findings emphasize the value of long-reads in resolving unrecognized genomic variation and highlight the importance of distinguishing biological from technical sequence differences.

Results and Discussion

P. brassicae has a T_4AG_3 telomere repeat composition and possess meiosis-related proteins in the nuclear genome.

By using the long-read SMRT technology, the numbers of nuclear assembled contigs were reduced from 165[16] to 20 (Supplementary Table S1). In the 25 Mb nuclear genome coding sequences were evenly distributed on each scaffold, interspersed with minor repeat regions (Fig. 1). Repetitive sequences were looked for manually and by using the tandem repeats finder[24]. Centromeric candidates with different length could be found on all nuclear contigs (Fig. 1). Thirteen of the 20 contigs of the nuclear genome (ranging from sizes between 692,149 bp to 2,120,846 bp) represent complete chromosomes from telomere to telomere (Fig. 1; Supplementary Table S2). The [TTTTAGGG] or T_4AG_3 telomeric sequences of *P. brassicae* (Supplementary Fig. S1) are identical with those found in *Chlamydomonas reinhardtii*, the green alga[25] and *Theileria annulata*, an apicomplexan animal parasite[26]. Among Archaeplastida, the T_3AG_3 telomeric motif is most common in plants for example in the *P. brassicae* e3 host *Arabidopsis thaliana* whereas the telomeres in red and green algae vary between T_4AG_3, T_3AG_3, T_2AG_3, and T_2AG_6[27,28]. Five of the remaining seven *P. brassicae* contigs were terminated by a single telomere. Efforts to further assemble all the remaining sequences into longer contigs terminated with telomeres were not successful. This could be explained by the enrichment of repetitive sequences at the ends making overlapping *in silico* analysis and additional PCR analysis followed by Sanger sequencing unsuccessful. Alternatively, the seven contigs may be dispensable or form supernumerary chromosomes. The organization of the nuclear chromosomes into two groups: a core set (permanent chromosomes) and accessory, dispensable, or supernumerary chromosomes is common among plant pathogenic fungi[29]. The extra chromosomes are known to play important roles for evolution, high recombination rates and adaptation to external changes[30]. Commonly the

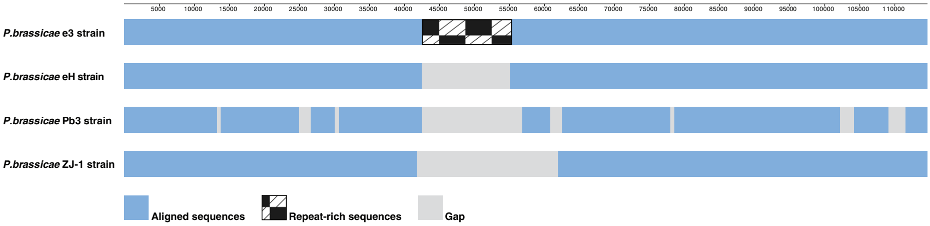


Figure 2. Alignment between four *Plasmodiophora brassicae* mitochondrial genomes. The illustration is based on mitochondrial genome synteny using the e3 strain (114,663 bp) as a template to which homologous regions identified in eH (102,962 bp), Pb3 (101,103 bp) and the ZJ-1 (93,640 bp) strain were aligned. For details, see Supplementary Fig. S9. The upper line represents sequence coordinates of the e3 mitochondrial genome.

dispensable chromosomes are enriched for genes involved in host colonization, infection and traits attributed to polymorphism observed between different isolates. Any such information and connection is yet not reported in *P. brassicae*. Earlier analyses based on pulse-field gel electrophoresis revealed between 6 to 16 chromosomal bands ranging from 680 kb to 2.2 Mb in size, including chromosomal polymorphisms between different single-spore and field isolates[12,31]. Whether the different observations reflect biological differences or are technical related remains to be clarified.

The MAKER[32] annotation pipeline combined with several *ab-initio* gene predictors generated 9,231 protein-encoding genes in the *P. brassicae* nuclear genome (Supplementary Table S1). In a comparison of *P. brassicae* with *B. natans, R. filosa* and the transcriptome of *S. subterranea* 5,605 proteins were earlier assigned to be *P. brassicae* specific[16]. In an extended and new comparison including additional transcriptome data from marine species[15] a joint core of 476 proteins shared with 12 other Rhizaria species was found. This revised analysis generated 3,017 *P. brassicae*-specific proteins whereof a majority lacked functional annotations (Supplementary Fig. S2; Supplementary Table S3).

Synaptonemal complexes (SC) were found by using serial thin sections of *P. brassicae* for electron microscopy[33]. SC, a proteinaceous structure, is known to assemble at the interface between pairs of homologous chromosomes at prophase I (zygotene) of meiosis I. SC formation is an essential feature of meiosis but is not strictly conserved in all organisms[34–36]. We used our new *P. brassicae* genome sequence to search for SC-associated and meiotic-related gene information (Supplementary Table S4) and next monitored their activity in enriched life stages of *P. brassicae* (Supplementary Fig. S3). Candidates such as *HOP1, ZIP1, ASY2, REC8, MER3, MSH5, FKBP6* were found to be highly active whereas animal orthologs such as C(3)G, SYCP1 and SYP2 were suppressed. SC proteins and important proteins for SC modification are known to vary on sequence level[36,37] which implies that additional genes with similar functions but not identified here can be present in *P. brassicae*. However, much remains to be learnt on SC components, their dynamics and regulation in general, and not least in *P. brassicae*.

P. brassicae has a large mitochondrial genome with a 12 kb repeat-rich region.

The mitochondrial genome represented by a single contig has a circular structure and a size of 114,663 bp (Supplementary Table S2). A uniform coverage with an average depth of ~800x was obtained across the genome, except from 47,000 to 50,000 bp where the depth decreased to ~400×. The possibility of miss-assembly was minimized by well-aligned reads spanning over this 3 kb stretch of AT-rich sequences that caused the reduced depth. The AT-rich sequences were found to be part of a 12,500 bp long repeat-rich region (42,650–55,150 bp) by a dot-plot self-similarity comparison (Supplementary Fig. S4a). A closer look into the dot plot indicated presence of tandem minisatellite- and microsatellite repeats with partially palindromic sequences in this region (Supplementary Fig. S4b).

Mt sequences are available for three *P. brassicae* strains: Pb3 from Canada[21], ZJ-1 from China[22] and eH from Germany[23]. In BLASTn search of the whole-genome shotgun database the mt sequence of e3 shared high identity (>99.95%) with these strains (Supplementary Table S5). However, the e3 mt sequence is about 11, 13 and 21 kb longer than eH, Pb3 and ZJ-1, respectively. We next compared the mt sequences from these three *P. brassicae* strains with e3, here visualized by dot-plots. While the sequences show high similarity for most of their length, the 12 kb repeat-rich region in e3 was not resolved in any other strain (Supplementary Figs. S5–S7). Neither was the repeat-rich region identified in the mt contigs we generated earlier by applying Illumina/454 technologies on the e3 genome[16]. The previously excluded sequence is provided here (Supplementary Table S6). The difference between the updated and former e3 mt sequence is visualized in Supplementary Fig. S8.

When analyzing mitochondrial synteny between the four *P. brassicae* strains, two locally collinear blocks (LCB) with highly conserved sequences and no rearrangements were identified by the whole genome alignment tool Mauve[38] (Supplementary Fig. S9). The area outside LCBs corresponds to the repeat-rich region identified in the e3 genome and lacked detectable homology in the other three genomes. If the homologous LCBs from eH, Pb3 and ZJ-1 are aligned to the e3 mt genome as illustrated in Fig. 2, absence of the repeat-rich region considerably contributes to genome size differences observed between the strains.

In conclusion, the mt intragenomic variation observed between the four strains (Fig. 2) is most likely technical rather than biological related. Several studies have reported unrecognized variation by short-read sequencing technologies and demonstrated that long reads which can span highly repetitive regions and thereby facilitate assembly, are essential for correct genome resolution[39,40].

The *P. brassicae* mt genome has a complex gene organization. We initially annotated the *P. brassicae* e3 mt sequence with MFannot, an automated tool commonly used for gene prediction in organelle genomes[41]. However, in the MFannot output many protein-coding and RNA genes were fragmented and with numerous introns predicted in intergenic regions, indicating interruption of coding sequences and incomplete annotations. Similar "mosaic" gene structure was reported in the eH mt genome annotated by MFannot[23]. To optimize *de novo* annotation, we combined automated predictions done with Prokka[42] and MAKER2[43] using Repeatmasker[44], tRNAscan-SE[45], Uniprot/Swiss-Prot mitochondrial proteins, the ribosomal database[46] and Rfam[47]. *P. brassicae* e3 transcriptome data[16] were re-assembled and mapped to the PacBio mt sequence to provide further information. The annotated mt genome of *S. subterranea*[48] was used as an additional source. All data were uploaded to Web Appollo[49], a web based annotation editing platform to facilitate manual curation.

This extensive annotation procedure identified seventy-nine genes in the *P. brassicae* e3 mt genome. Thirty-two are predicted as protein-coding, 28 as structural RNA genes and 19 are ORFs (Fig. 3a; Supplementary Table S7). Seventeen protein-coding genes are involved in the mitochondrial respiratory chain, ten genes code for small ribosomal subunit proteins, four for proteins of the large ribosomal subunit and one gene (*rdp*) is coding for a RNA-directed DNA polymerase (Supplementary Table S8). Structural RNAs include large and small ribosomal subunits (*rnl* and *rns*), 5S ribosomal RNA (*rrn5*), a ribonuclease P type B (*rnpB*) and 24 transfer RNAs (Supplementary Table S9). By sequence similarity searches of the Uniprot/Swiss-Prot database (e-value < 10e-6) functional information was retrieved for ten ORFs (3 intergenic and 7 within introns) while nine ORFs (4 intergenic and 5 within introns) had no significant similarity (Supplementary Table S10). Seven intron group II and 2 transposon-like elements (Supplementary Table S11) and few simple repeats (Supplementary Table S12) were also found in the e3 mt genome.

In the *P. brassicae* eH strain sixty mt genes were identified[23], divided on 32 protein-coding and 28 structural RNA genes corresponding to numbers and functional categories in e3. No transposable elements, ORFs, intron group II or repeats were reported. A detailed comparison of translated coding sequences could not be carried out here because amino acid sequences are not available for eH.

In comparison to its closely related potato scab pathogen *S. subterranea*[48], the *P. brassicae* mt genome is roughly three times larger (Fig. 3a–f; Supplementary Table S7). When aligned, the genomes shared two syntenic blocks with well-conserved sequences, free from rearrangements (Fig. 4). Homology with the *S. subterranea* mt genome was not found in the 12 kb repeat-rich region in *P. brassicae* e3. Otherwise, the gene order between the genomes is nearly identical as shown in the gene maps (Fig. 3a,f). The map view demonstrates that presence of the repeat-rich region, high numbers of introns and variation in intergenic regions contributed to larger size of the *P. brassicae* e3 mt genome. Due to a larger mt genome size, *P. brassicae* e3 has a slightly lower coding density (54%) than *S. subterranea* (66%) (Supplementary Table S7).

When comparing the incidence of protein-coding genes, three genes (*atp8*, *rpl5* and *rps10*) were missing in the *S. subterranea* mt sequence (Supplementary Table S8). Several small ribosomal subunit proteins share overlapping sequences in the e3 mt genome, including *rps7/rps12* (19 bp), *rps11/rps13* (30 bp) and *rps3/rps19* (22 bp). Similar sequence overlaps between ribosomal genes are also seen in the *S. subterranea* mt genome[48]. The majority of e3 protein-coding genes (23 out of 32) start with the standard ATG codon (methionine) whereas TTA and TTG (leucine) seem to be used by five genes, and ATC and ATT (isoleucine) by three genes as alternative initiation codons (Supplementary Table S13). The GTG codon (valine) is employed only by *nad5*, in contrast to its extensive use among mt proteins in *Lotharella oceanica*, the rhizarian chlorarachniophyte alga[50]. Most of the e3 genes are terminated by TAA and in few cases with TAG codon. TGA, the third standard stop codon is translated into tryptophan in the mt genomes of *P. brassicae* e3 and *S. subterranea*[48]. This TGA usage pattern is therefore important to consider when selecting a correct genetic code for translation of *P. brassicae* mt coding sequences.

All structural RNA genes except arginine (*trnR*), proline (*trnP*) and tryptophan (*trnW*) were encoded on an 11,332 bp segment in the *P. brassicae* e3 mt genome (Fig. 3a). Densely located ribosomal RNA genes have been found in several other mt genomes and are believed to be involved in balanced co-transcription and RNA turnover[51]. Total number of tRNAs (24) was identical to that found in *S. subterranea*, with the difference that the e3 mt genome codes for glutamine (*trnQ*) and one copy of histidine (*trnH*) (Supplementary Table S9).

In contrast to *S. subterranea*, the *P. brassicae* e3 mt genome is strikingly intron-rich. Twelve genes in the e3 mitochondrial respiratory chain harbor the majority (41 out of 54) of the introns (Supplementary Table S8) while the remaining 13 introns are found in the large and small ribosomal subunit (Supplementary Table S9). Only five introns in three mt protein-coding genes were found in *S. subterranea*[48], whereas chlorarachniophyte algae, *B. natans* and *L. oceanica* lack introns in their mt genomes[50]. Defining exon-intron splicing sites, especially in the genes with multiple exons was a challenge during annotation of the e3 mt genome and some positions remain to be exactly clarified. Application of the canonical acceptor-donor sites (GT-AG, GC-AG, AT-AC) resulted in reading frame shifts and disruptions of coding sequences predicted based on sequence alignments with known proteins. According to our current annotations supported by transcriptome data[16], a new splicing pattern with the GT-AT and GA-CA acceptor-donor sites is likely employed by the *P. brassicae* e3 mt genome. Further, seven group II introns and two transposon-like elements, including Long Terminal Repeat (*LTR*) and Enhancer/Suppressor-mutator (*En/Spm*) were identified in the e3 mt genome (Fig. 3a; Supplementary Table S11). All these sequences except *En/Spm* were found to overlap exon-intron boundaries and may have an important role in the splicing mechanism. The MFannot tool[41] predicted much higher number of group II introns (45) whereof 34 in gene regions (Supplementary Table S14). These introns were not submitted to the European Nucleotide Archive since no other annotation software supported their prediction. Annotated mt sequences from other plasmodiophorids are needed to resolve whether the higher number of group II introns represents an overestimated or a common landmark. Out of nineteen ORFs predicted in the *P. brassicae* e3 mt genome, the majority (12) was found within introns of the protein-coding genes (Supplementary Table S7). Seven of these ORFs showed significant similarity (e-value < 10e-6) with known proteins including among others three maturase-like proteins (*mat*1 and

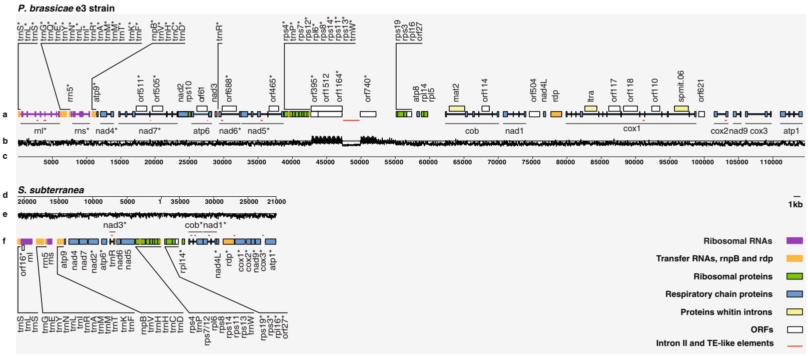

P. brassicae e3 strain
trnS* trnL* trnS trnG* trnQ* trnE* trnY* trnN* trnL* trnI* trnR* trnA* trnM* trnM* trnT* trnK* trnF*
rnpB* trnV* trnH* trnC* trnD*
trnR*
rps4* trnP* rps7* rps12* rpl6* rps8* rps14* rps11* rps13* trnW*
rps19 rps3 rpl16 orf27
rrn5* atp9* orf511* orf505* nad2 rps10 orf61 nad3 orf688* orf465* orf395* orf1512 orf1164* orf740* atp8 rpl14 rpl5 mat2 orf114 orf504 nad4L rdp ltra orf117 orf118 orf110 spmit.06 orf621
a
rnl* rns* nad4* nad7* atp6 nad6* nad5* cob nad1 cox1 cox2 nad9 cox3 atp1
b
c
5000 10000 15000 20000 25000 30000 35000 40000 45000 50000 55000 60000 65000 70000 75000 80000 85000 90000 95000 100000 105000 110000
S. subterranea
d
20000 15000 10000 5000 1 35000 30000 25000 21000
1kb
e
nad3* cob* nad1*
f
orf16* rnl rrn5 rns atp9 nad4 nad7 nad2* atp6* trnR nad6 nad5 rpl14* nad4L* rdp* cox1* cox2* nad9* cox3* atp1*
trnS trnL trnS trnG trnE trnY trnN trnL trnI trnR trnA trnM trnM trnT trnK trnF rnpB trnV trnH trnH trnC trnD rps4 trnP rps7/12 rpl6 rps8 rps14 rps11 rps13 trnW rps19* rps3* rpl16* orf27*
Ribosomal RNAs
Transfer RNAs, rnpB and rdp
Ribosomal proteins
Respiratory chain proteins
Proteins whitin introns
ORFs
Intron II and TE-like elements

Figure 3. Physical maps and GC content of the *Plasmodiophora brassicae* e3 and *Spongospora subterranea* mitochondrial genomes. The two mitochondrial genomes have circular structure but are here linearized to facilitate comparisons. Boxes represent genes and exons separated by introns illustrated as lines. Colors correspond to specified gene functional groups. Genes transcribed from right to left have names marked with an asterisk. (**a**) Physical map of the *P. brassicae* e3 mitochondrial genome. (**b**) GC content of the *P. brassicae* e3 mitochondrial genome in non-overlapping sliding 20 bp windows. The horizontal line represents 50% GC. (**c**) Sequence coordinates of the *P. brassicae* e3 mitochondrial genome. (**d**) Sequence coordinates of the *S. subterranea* mitochondrial genome. For comparison, the *S. subterranea* sequence was opened at position 21,000 bp and two fragments (21,001-1 bp and 37,699–21,000 bp) were merged. (**e**) GC content of the *S. subterranea* mitochondrial genome in non-overlapping sliding 20 bp windows. The horizontal line represents 50% GC. (**f**) Physical map of the *S. subterranea* mitochondrial genome.

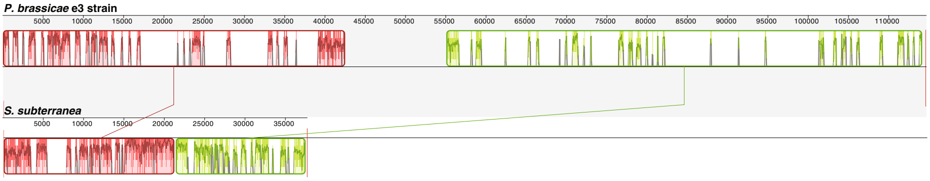


Figure 4. Synteny between the *Plasmodiophora brassicae* e3 and *Spongospora subterranea* mitochondrial genomes. The image[38] illustrates aligned mitochondrial sequences of *P. brassicae* e3 (upper panel) and *S. subterranea* (lower panel). Identified homologous regions are displayed as two locally collinear blocks (LCB), green and red. The LCBs with matching colors between the genomes are connected by lines. Inside each LCB, a sequence similarity profile is shown, with its height representing the average level of sequence conservation. White areas within a LCB indicate sequences that are unique to a particular genome. The region outside LCBs (*P. brassicae* e3 sequence from 42,547 to 55,095 bp) has no detectable homology with the *S. subterranea* mitochondrial genome.

*mat*2), a group II intron-encoded protein (*LtrA*) and a DNA binding endonuclease (*spmit*.06) (Supplementary Table S10). Much remains to understand about evolution and regulatory function of the complex intron pattern in the *P. brassicae* e3 mt genome. It is intriguing to speculate that intron-encoded maturases could assist in the splicing processes, maybe promoting self-splicing as observed in fungal mitochondria[52].

Overall GC content is close to identical in the mt genomes of *P. brassicae* e3 (26.2%) and *S. subterranea* (26.8%)[48] while being lower compared to the chlorarachniophyte algae *B. natans* (42.2%) and *L. oceanica* (50.1%)[50]. However, large variation in the GC content was observed across the 12 kb repeat-rich region in *P. brassicae* e3 (Fig. 3b). A drastic increase above 75% of GC content was present from 42,800 to 47,200 bp, in the region where tandem minisatellites were identified (Supplementary Fig. S4b). Two ORFs coding for hypothetical proteins and one ORF with high similarity to thrombospondin motifs were annotated in this region (Fig. 3a; Supplementary Table S10). A sharp drop to below 25% of GC content was observed from 47,300 to 49,800 bp, corresponding to the AT-rich stretch of tandem microsatellite repeats (Supplementary Fig. S4b). This partially palindromic 2.5 kb sequence was annotated as a *En/Spm* transposon-like element (Fig. 3a; Supplementary Table S11). A second increase of GC content (>75%) occurred from 49,900 to 52,150 bp, in the region with tandem minisatellites (Supplementary Fig. S4b), and an ORF with high similarity to a proline-rich protein (Fig. 3a; Supplementary Table S10). The GC content varied between 75% and 50% from 52,200 to 55,000 bp, where minisatellites were detected on the end of repeat-rich region (Supplementary Fig. S4b). Distinct switches in GC composition are used

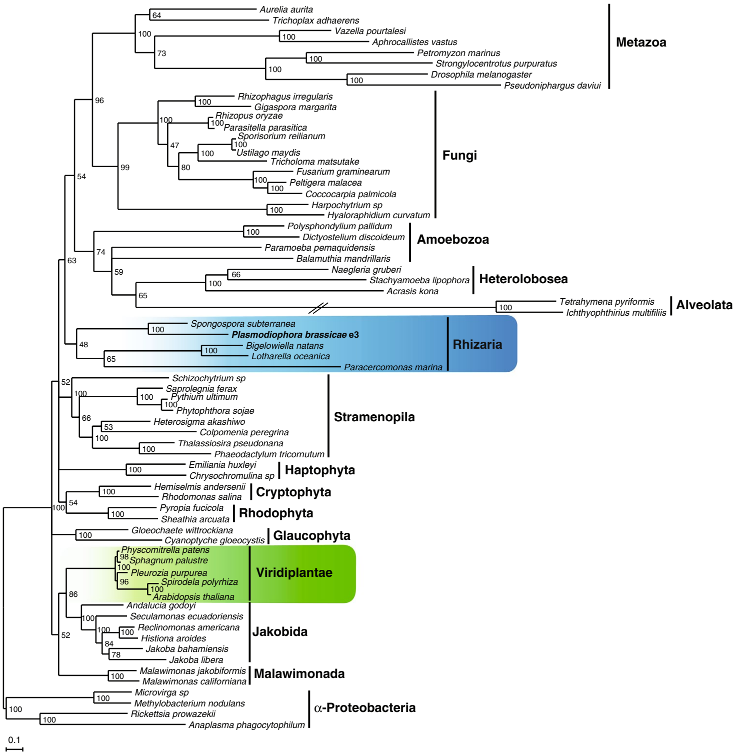


Figure 5. Maximum likelihood tree inferred from mitochondrial-encoded proteins. The tree was constructed from a concatenated alignment of 12 mitochondrial protein sequences from 67 organisms (63 from major eukaryotic groups and 4 α-protebacteria used as outgroup species). The super-matrix was analyzed with RAxML 8.2.11[60] using GAMMA and substitution models specified for each partition. For details, see Supplementary Table S15. Rapid bootstrap analysis was done with 250 iterations. Branches with support values <45 were collapsed. The scale bar shows the inferred number of amino acid substitutions per site. The long branch was reduced to 50% of its original length, indicated by a crossed double line.

as genomic signatures of possible horizontal or lateral gene transfers particularly in bacteria[53,54]. Whether the observed fluctuations in GC content in the *P. brassicae*-specific mitochondrial region 42,800 to 55,000 bp region could be a remnant of ancient genomic events remain to be elucidated.

The SAR group (Stramenopila, Alveolata, Rhizaria) comprises numerous diverse eukaryotic organisms on which revised phylogenetic relationships have been presented[2]. We attempted to generate new information based on the *P. brassicae* e3 mt genome, since no phylogeny is yet proposed for the Phytomyxea plant pathogens based on a larger set of mitochondrial genes. Amino acid sequences for 12 protein-coding genes (*cob*, *cox1*, *cox2*, *cox3*, *atp6*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L, nad5* and *nad6*) from 67 organisms were used to infer maximum likelihood RAxML single-gene and concatenated trees. Twelve single-gene trees showed a well-supported close relationship between *P. brassicae* and *S. subterranea* as well as between *B. natans* and *L. oceanica*. When 773 mt proteins were analyzed, the five species in Rhizaria clustered together and *P. brassicae* and *S. subterranea* remained the two most closely related plasmodiophorids in the concatenated tree (Fig. 5). Moderate support for Rhizaria and several other major clades most likely reflects limitations in sequence information. However, no evidence for horizontal gene transfer events to *P. brassicae* neither from any species in Viridiplantae including *Arabidopsis thaliana*, which can act as a host, nor from other organisms was detected.

Conclusion

Our long-read sequencing approach provided new insights into the nuclear and mitochondrial genomes of the clubroot pathogen *P. brassicae*. For the first time telomeric sequences are presented in the supergroup Rhizaria. Whether *P. brassicae* and algae share evolutionary history is intriguing. Both the telomeric T_4AG_3 repeats and some mitochondrial ORFs in *P. brassicae* showed high sequence similarities with algae and point to that direction.

The close relationship between *P. brassicae* and *S. subterranea* was further visualized upon mt genome comparisons. A number of peculiarities were found in the *P. brassicae* mt genome such as: a unique repeat-rich region, a new splicing pattern with the GT-AT and GA-CA acceptor-donor sites, group II introns spanning exon-intron boundaries, and sequence similarities in ORFs to functional genes present in various organisms.

Methods

Materials and sequencing.

Resting spores from clubs of *Brassica rapa* grown in *Plasmodiophora brassicae* strain e3 infested soils were isolated and used for DNA extraction[55]. High-quality DNA was sent to SciLifeLab, Uppsala, Sweden for PacBio RSII sequencing according to the manufacturer's protocol.

Nuclear assembly and annotation.

Raw reads were assembled using FALCON v0.4 and HAGP3 from SMRTportal v2.3. The two assemblies were manually merged and polished using Quiver from SMRTportal v2.3. The gene annotation pipeline MAKER v2.3[32] was used in combination with *ab-initio* gene predictors: Augustus v2.5.5, SNAP and GeneMark-ES v2.3. Augusts and SNAP were trained on the previously annotated and manually curated *P. brassicae* nuclear genes[16]. The UniProt/Swiss-Prot database and all rhizarian ESTs and proteins found at NCBI, as well as transcripts assembled from strand-specific RNASeq *P. brassicae* libraries[16] were used as evidence integrated in gene predictions. Annotation of repeats was performed within MAKER[32], using a *P. brassicae* specific repeat library constructed *de novo* using RepeatModeler v1.0.7 as well as MAKER's internal library of transposable elements and the Repbase repeat library rm-20130422. MAKER was run using default parameters except *pred_flank* that was set to 100 bp, *split_hit*. Telomeric sequences were identified by using Tandem Repeats Finder[24]. For biological pathway classifications we used the WebMGA and the KOG classification tools. Additional protein analysis was done using OrthoMCL, sequence similarity searches, BLASTP searches against GenBank non-redundant protein database, HMM-searches against Pfam database, and RPS-BLAST searches against NCBI KOG (March 2017).

Mitochondrial genome *de novo* assembly and annotation.

The contig encoding the mitochondrial genome was assembled using Canu v1.5[56]. Visualization of the raw assembly with Bandage generated a circular contig (133,222 bp) in which overlapping sequences were identified by Gepard v1.40[57]. After removing overlaps (18,559 bp) and circularization, the final 114,663 bp long PacBio sequence was polished using Quiver. Mapping of the Illumina data[16] to the PacBio sequence using BWA v0.7.15 and SAMtools v1.5 and polishing with Pilon revealed 2 × 1 bp difference, which were corrected. Next, the coverage was tested by aligning the PacBio reads to the assembled mitochondrial genome using GraphMap v0.5.2[58].

To optimize *de novo* annotation several tools and sources were used and combined with manual curation. Automated annotations were done using MFannot v1.33[41] with the genetic code 4 "Mold, protozoan and coelenterate mitochondrial; Mycoplasma/Spiroplasma", Prokka v1.1[42] with mitochondria and archaea kingdoms and MAKER2[43] using Repeatmasker[44], tRNAscan-SE[45], Uniprot/Swiss-Prot mitochondrial proteins (Nov. 2016), the ribosomal database[46] and Rfam v12[47]. Further information was provided by transcriptome data[16], re-assembled using Trimmomatic, Tophat and Stringtie. The *S. subterranea* annotated mitochondrial genome[48] was used as an additional source. Based on the different lines of annotation and sources, the gene models have been manually created through Web Apollo[49]. Translated CDS features were blasted against the Uniprot/Swiss-Prot reference data set (Aug. 2016) and filtered using a maximal e-value of 10e-6 and run against InterProScan v5.21–60. All retrieved functional information have been integrated into the final annotated data set. Predicted ORFs were used as query sequences for sequence similarity searches of the Uniprot/Swiss-Prot database (March 2019).

Mitochondrial sequence comparison.

Dot-plots created by Gepard v1.40[57] were used for comparison of *P. brassicae* mitochondrial sequences from four strains: e3 (generated in this study, GeneBank accession LS992577), eH[23] (GeneBank accession POCA01000043), ZJ-1[22] (GeneBank accession MCBL01000050) and the Pb3 strain[21] (GeneBank accession RZOB01000060). Additional sequence comparison was done using two mitochondrial contigs of the e3 strain generated by Illumina/454, excluded from[16] but provided in Supplementary Table S6. Synteny between mitochondrial sequences from four *P. brassicae* strains and *S. subterranea*[48] (GeneBank accession KF738139) was tested using the Mauve genome alignment tool v2.4.0[38] with default settings.

Phylogenetic analysis.

Amino acid sequences were retrieved from public databases for 12 mitochondrial protein-coding genes (*cob*, *cox1*, *cox2*, *cox3*, *atp6*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L, nad5* and *nad6*) conserved across 67 organisms. Sixty-three organisms were selected to represent major eukaryotic groups with available complete mitochondrial genomes and if possible, deep-branching positions. Four α-protebacteria were selected as outgroup species. The sequences were aligned using Clustal Omega v1.2.1[59] with default settings. Multiple alignments were visualized with AliView, examined and automatically trimmed with trimAL v1.4. Maximum likelihood (ML) phylogenetic trees with rapid bootstrap (RB) analyses were generated with RAxML v8.2.11[60]. The best amino acid substitution model was estimated with PROTGAMMAUTO option for each single gene tree and run with 250 to 650 RB, a number of iterations predicted to be sufficient by the autoFC stopping criteria. For concatenated trees, 12 protein alignments were concatenated (the script is available at https://github.com/nylander/catfasta2phyml) into a super-matrix comprising 773 genes and 3,819 aligned amino acid positions. ML analyses were inferred under GAMMA rate of heterogeneity with the substitution models specified for each partition and run with 250 RB iterations. Trees were displayed and edited with Dendroscope v3.5.9[61]. Information on organisms and proteins and substitution models are listed in Supplementary Table S15.

Data availability

Data retrieved in this study are deposited in the European Nucleotide Archive under the project PRJEB24736. For the nuclear sequence under accession number OVEO01000001-OVEO01000020 and the mt sequence under accession number LS992577.

Received: 17 June 2019; Accepted: 15 October 2019;
Published online: 31 October 2019

References

1. Foissner, W. & Hawksworth, E. *Protist diversity and geographical distribution* (eds Foissner, W & Hawksworth D). Vol. 8. Topics in biodiversity and conservation. (Springer, 2009).
2. Cavalier-Smith, T., Chao, E. E. & Lewis, R. Multigene phylogeny and cell evolution of chromist infrakingdom Rhizaria: contrasting cell organization of sister phyla Cercocoa and Retaria. *Protoplasma* **255**, 1517–1574 (2018).
3. Neuhauser, S., Kirchmair, M. & Gleason, F. H. Ecological roles of the parasitic phytomyxids (plasmodiophorids) in marine ecosystems: a review. *Marine & Freshwater Res.* **62**, 365–371 (2011).
4. Dixon, G. R. The occurrence and economic impact of *Plasmodiophora brassicae* and clubroot disease. *J. Plant Growth Regul.* **28**, 194–202 (2009).
5. Wallenhammar, A.-C. *et al.* Clubroot, a persistent threat in Swedish oilseed rape production. *Can. J. Plant Pathol.* **36**, 135–141 (2014).
6. Botero, A. *et al*. Clubroot disease in Latin America: distribution and management strategies. *Plant Pathol.* **68**, 827–833 (2019).
7. Ingram, D. S. & Tommerup, I. C. The life history of *Plasmodiophora brassicae* Woron. *Proc. R. Soc. Lond. B.* **180**, 103–112 (1972).
8. McDonald, M. R. *et al*. The role of primary and secondary infection in host response to *Plasmodiophora brassicae*. *Phytopathol.* **104**, 1078–1087 (2014).
9. Kageyma, K. & Asano, T. Life cycle of *Plasmodiophora brassicae*. *J. Plant Growth Regul.* **28**, 203–211 (2009).
10. Siemens, J., Bulman, S., Rehn, F. & Sundelin, T. Molecular biology of *Plasmodiophora brassicae*. *J. Plant Growth Regul.* **28**, 245–251 (2009).
11. Schwelm, A., Dixelius, C. & Ludwig-Müller, J. New kid on the block – the clubroot pathogen genome moves the plasmodiophorids into the genomic era. *Eur. J. Plant Pathol.* **145**, 531–542 (2016).
12. Strelkov, S. E. *et al*. Virulence and pathotype classification of *Plasmodiophora brassicae* populations collected from clubroot resistant canola (*Brassica napus*) in Canada. *Can. J. Plant Pathol.* **40**, 284–298 (2018).
13. Curtis, B. A. *et al*. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**, 59–65 (2012).
14. Glöckner, G. *et al. The genome of the foraminiferan Reticulomyxa filosa. Curr. Biol.* **24**, 11–18 (2014).
15. Keeling, P. J. *et al.* The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the function a diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12**, e1001889 (2014).
16. Schwelm, A. *et al*. The *Plasmodiophora brassicae* genome reveals insights in its life cycle and ancestry of chitin synthases. *Sci. Rep.* **5**, 11153 (2015).
17. Krabberød, A. K. *et al*. Single cell transcriptomics, mega-phylogeny, and the genetic basis of morphological innovations in Rhizaria. *Mol. Biol. & Evol.* **34**, 1557–1573 (2017).
18. Ciaghi, S., Neuhauser, S. & Schwelm, A. Draft genome resource for the potato powdery scab pathogen *Spongospora subterranea*. *Mol. Plant-Microbe Interact.* **31**, 1227–1229 (2018).
19. Falloon, R. E. *et al*. Root infection of potato by *Spongospora subterranea*: knowledge review and evidence for decreased plant productivity. *Plant Pathol.* **65**, 422–434 (2016).
20. Sierra, R. *et al*. Evolutionary origins of Rhizarian parasites. *Mol. Biol. Evol.* **33**, 980–983 (2016).
21. Rolfe, S. A. *et al*. The compact genome of the plant pathogen *Plasmodiophora brassicae* is adapted to intracellular interactions with host *Brassica* spp. *BMC Genomics* **17**, 1–15 (2016).
22. Bi, K. *et al*. Integrated omics study of lipid droplets from *Plasmodiophora brassicae*. *Scientific Rep.* **6**, 36965 (2016).
23. Daval, S. *et al*. Computational analysis of the *Plasmodiophora brassicae* genome: mitochondrial sequence description and metabolic pathway database design. *Genomics* https://doi.org/10.1016/j.ygeno.2018.11.013
24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
25. Petracek, M. E., Lefebvre, P. A., Silflow, C. D. & Berman, J. Chlamydomonas telomere sequences are A+T-rich but contain three consecutive G-C base pairs. *Proc. Nat. Acad. Sci. USA* **87**, 8222–8226 (1990).
26. Sohanpal, B. K., Morzaria, S. P., Gobright, E. I. & Bishop, R. P. Characterisation of the telomeres at opposite ends of a 3 Mb *Theileria parva* chromosome. *Nucleic Acids Res.* **23**, 1942–1947 (1995).
27. Fredrychova, R. C. & Mason, J. M. Telomeres: their structure and maintenance. *In: Stuart D., ed. Mechanisms of DNA replication. INTECH Open Science* **17**, 423–443 (2013).
28. Procházková Schrumpfová, P., Schořová, Š. & Fajkus, J. Telomere- and telomerase-associated proteins and their functions in the plant cell. *Front. Plant Sci.* **7**, 851 (2016).
29. Mehrabi, R., Gohari, A. M. & Kema, G. H. J. Karyotype variability in plant-pathogenic fungi. *Ann. Rev. Phytopathol.* **55**, 483–503 (2017).
30. Croll, D. & McDonald, B. A. The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathog.* **8**, e1002608 (2012).
31. Strehlow, B., de Mol, F. & Struck, C. History of oilseed rape cropping and geographic origin affect the genetic structure of *Plasmodiophora brassicae* populations. *Phytopathol.* **104**, 532–538 (2014).
32. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
33. Braselton, J. P. Karyotypic analysis of *Plasmodiophora brassicae* based on serial thin-sections of pachytene nuclei. *Can. J. Bot.* **60**, 403–408 (1982).
34. Giroux, C. N., Dresser, M. E. & Tiano, H. F. Genetic control of chromosome synapsis in yeast meiosis. *Genome* **31**, 88–94 (1989).
35. Qiao, H. *et al*. Interplay between synaptonemal complex, homologous recombination, and centromeres during mammalian meiosis. *PLoS Genet.* **8**, e1002790 (2012).
36. Grishaeva, T. M. & Bogdanov, Y. F. Conservation and variability of synaptonemal complex proteins in phylogenesis of eukaryotes. *Int. J. Evol. Biol.* **2014**, 856230 (2014).
37. Gao, J. & Colaiácova, M. P. Zipping and unzipping: protein modifications regulating synaptonemal complex dynamics. *Trends Genet* **34**, 232–245 (2018).
38. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
39. Oren, M. *et al*. Short tandem repeats, segmental duplications, gene deletion, and genomic instability in a rapidly diversified immune gene family. *BMC Genom.* **17**, 900 (2016).
40. Paajanen, P. *et al*. A critical comparison of technologies for a plant genome sequencing project. *Gigascience* **8**, 1–12 (2019).

41. Beck, N. & Lang, B.F. MFannot, organelle genome annotation webserver. http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl (2010).
42. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
43. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
44. Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker Open-3.0. http://www.repeatmasker.org (2010).
45. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoScan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33**, 686–689 (2005).
46. Cole, J. R. *et al*. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, 633–642 (2014).
47. Nawrocki, E. P. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–137 (2015).
48. Gutiérrez, P., Bulman, S., Alzate, J., Ortíz, M. C. & Marin, M. Mitochondrial genome sequence of the potato powdery scab pathogen *Spongospora subterranea*. *Mitochondrial DNA Part A* **27**, 58–59 (2016).
49. Lee, E. *et al*. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.* **14**, R93 (2013).
50. Tanifuji, G., Archibald, J. M. & Hashimoto, T. Comparative genomics of mitochondria in chlorarachniophyte algae: endosymbiotic gene transfer and organellar genome dynamics. *Sci. Rep.* **6**, 21016 (2016).
51. Valach, M., Moreira, S., Kiethega, G. N. & Burger, G. Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria. *Nucleic Acids Res.* **42**, 2660–2672 (2014).
52. Guha, T. K., Wai, A., Mullineux, S.-T. & Hausner, G. The intron landscape of the mtDNA *cytb* gene among the Ascomycota: introns and intron-encoded open reading frames. *Mitochondrial DNA Part A* **29**, 1015–1024 (2018).
53. Hayek, N. Lateral transfer and GC content of bacterial resistance genes. *Front. Microbiol.* **4**, 41 (2013).
54. Zhang, D. *et al*. Root parasitic plant *Orobanche aegyptiaca* and shoot parasitic plant *Cuscuta australis* obtained Brassicaceae-specific strictosidine synthase-like genes by horizontal gene transfer. *BMC Plant Biol.* **14**, 19 (2014).
55. Mehrabi, S., Stjelja, S. & Dixelius, C. Disease establishment, resting spore isolation and DNA extraction of *Plasmodiophora brassicae*, the clubroot pathogen. *Bio-Protocol* **101**, e2864 (2018).
56. Koren, S. *et al*. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
57. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
58. Sović, I. *et al*. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Com.* **7**, 11307 (2016).
59. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Systems Biol.* **7**, 539 (2011).
60. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
61. Huson, D. & Scornavacca, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *System. Biol.* **61**, 1061–1067 (2012).

Acknowledgements

We thank Jacques Dainat and Mahesh Binzer-Panchal at the National Bioinformatics Infrastructure Sweden (NBIS) for extra support on mitochondrial genome bioinformatics. This work was funded by the Swedish Research Council FORMAS, and the Swedish University of Agricultural Sciences (SLU). The Science for Life Laboratory and the Swedish National Genomics Infrastructure (NGI) were supported from the Knut and Alice Wallenberg Foundation and the Swedish research council VR. Open access funding provided by Swedish University of Agricultural Sciences.

Author contributions

S.S., J.F. and C.D. planned and designed the research. S.S. performed experiments and S.S., J.F. and C.T.-R. analyzed data. S.S., J.F. and C.D. wrote the manuscript, with contribution and revision by C.T.-R.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-52274-7.

Correspondence and requests for materials should be addressed to C.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2021:70

This thesis enhances the genomic knowledge on *Plasmodiophora brassicae*, the causative agent of clubroot. Based on long-read sequences *de-novo* genome assembly was generated. A nuclear genome (25.2 Mb) was represented with 20 contigs, characterized by resolution of repetitive regions, identification of telomeres, prediction of 9,231 protein-coding genes and selection of effector candidates. A complex, intron-rich sequence organization was revealed in a 114.6 kb mitochondrial genome. These findings and genome sequences represent a valuable resource for *P. brassicae* future studies.

Suzana Stjelja Arvelius received her graduate education at the Department of Plant Biology, SLU, Uppsala. She received her MSc in Animal Breeding and Genetics (EM-ABG) at SLU and UMB, Norway and her BSc in Animal Science from the University of Belgrade, Serbia.

Online publication of thesis summary: http://pub.epsilon.slu.se/

ISSN 1652-6880
ISBN (print version) 978-91-7760-817-2
ISBN (electronic version) 978-91-7760-818-9