PacBio Hi-Fi genome assembly of Sipha maydis, a model for the study of multipartite mutualism in insects

François Renoz, Nicolas Parisot, Patrice Baa-Puyoulet, Léo Gerlin, Samir Fakhour, Hubert Charles, Thierry Hance, Federica Calevro

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

PacBio Hi-Fi genome assembly of *Sipha maydis*, a model for the study of multipartite mutualism in insects

François Renoz[1,2,3,6]✉, Nicolas Parisot[2,6]✉, Patrice Baa-Puyoulet[4], Léo Gerlin[4], Samir Fakhour[1,5], Hubert Charles[2], Thierry Hance[1] & Federica Calevro[4]✉

Dependence on multiple nutritional endosymbionts has evolved repeatedly in insects feeding on unbalanced diets. However, reference genomes for species hosting multi-symbiotic nutritional systems are lacking, even though they are essential for deciphering the processes governing cooperative life between insects and anatomically integrated symbionts. The cereal aphid *Sipha maydis* is a promising model for addressing these issues, as it has evolved a nutritional dependence on two bacterial endosymbionts that complement each other. In this study, we used PacBio High fidelity (HiFi) long-read sequencing to generate a highly contiguous genome assembly of *S. maydis* with a length of 410 Mb, 3,570 contigs with a contig N50 length of 187 kb, and BUSCO completeness of 95.5%. We identified 117 Mb of repetitive sequences, accounting for 29% of the genome assembly, and predicted 24,453 protein-coding genes, of which 2,541 were predicted enzymes included in an integrated metabolic network with the two aphid-associated endosymbionts. These resources provide valuable genetic and metabolic information for understanding the evolution and functioning of multi-symbiotic systems in insects.

Background & Summary

Nutritional symbiosis with bacteria has contributed significantly to the evolutionary success of insect taxa that feed on unbalanced diets such as phloem sap, blood or wood[1]. Indeed, in many insect species, the synthesis of nutrients (e.g. amino acids and/or vitamins) that are not present in sufficient amounts in the diet is ensured by an obligate nutritional symbiont, sometimes acquired tens of millions of years ago[2]. These symbionts are generally transmitted faithfully from generation to generation (i.e., by vertical transmission), and compartmentalized in specific host cells called bacteriocytes[3]. These cells mediate the metabolic exchanges between the insect and its bacterial partners, and regulate populations of obligate symbionts according to the insect's nutritional needs throughout its life cycle[4,5]. However, this intracellular lifestyle causes Muller's ratchet which, combined with severe population bottlenecks during vertical transmission and the relaxation of purifying selection on genes no longer needed in the context of interdependent association, leads to reductive genome evolution[6]. In some insect lineages, the ancestral nutritional symbiont has undergone such severe genomic erosion that it is no longer able to supply alone the compounds essential to its host's physiology on its own, and is metabolically complemented by more recently acquired nutritional symbionts[1,7]. Thus, in many insect taxa, nutritional symbiosis does not rely on a single obligate symbiont but on a consortium of nutritional symbionts that evolve together in the same host, either within the same bacteriocytes[8–11], or in distinct but anatomically connected bacteriocytes[12–16].

These multi-partner symbiotic system have evolved in a wide range of hemipteran taxa, including several insect pests, including Psylloidae (psyllids[13,17–19]), Aleyrodidae (whiteflies[20–23]), Pseudococcinae (mealybugs[8,24–26]), Auchenorrhyncha (cicadas, leafhoppers, planthoppers and treehoppers[12,27–29]), Adelgidae (adelgids[15,30–35]) and Aphidoidea (aphids[14,36–41]). However, little is known about the development and functioning of these systems. This is largely due to the fact that their study is hampered by a range of constraints such as

[1]Biodiversity Research Centre, Earth and Life Institute, UCLouvain, Louvain-la-Neuve, 1348, Belgium. [2]Univ Lyon, INSA Lyon, INRAE, BF2I, UMR203, Villeurbanne, F-69621, France. [3]Institute of Agrobiological Sciences, National Agriculture and Food Research Organization (NARO), Tsukuba, Ibaraki, 305-8634, Japan. [4]Univ Lyon, INRAE, INSA Lyon, BF2I, UMR203, Villeurbanne, F-69621, France. [5]Department of Plant Protection, National Institute for Agricultural Research (INRA), Béni-Mellal, 23000, Morocco. [6]These authors contributed equally: François Renoz, Nicolas Parisot. ✉e-mail: francois.renoz@uclouvain.be; nicolas.parisot@insa-lyon.fr; federica.calevro@insa-lyon.fr

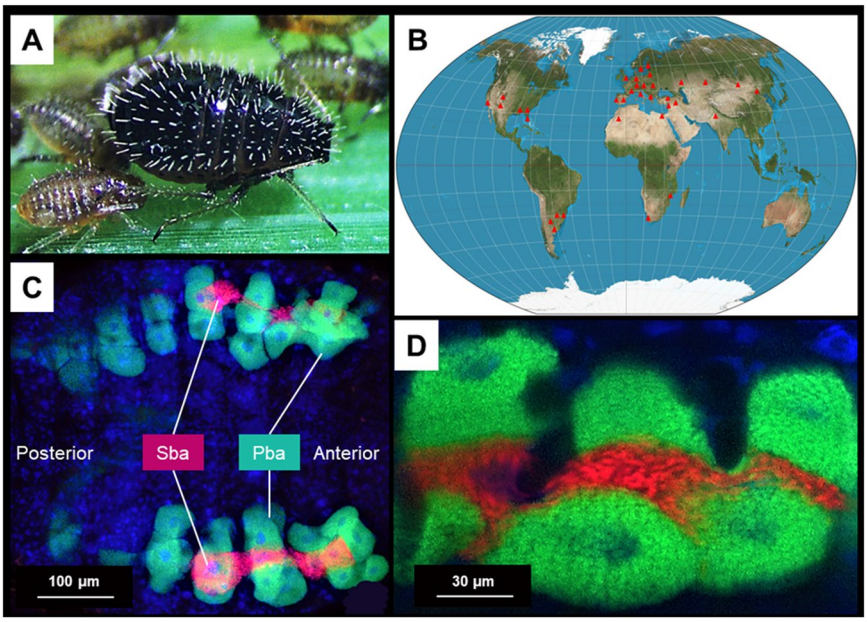

Fig. 1 The cereal aphid *Sipha maydis* and its di-symbiotic system. (**A**) An adult surrounded by several nymphs, all feeding on bread wheat, *Triticum aestivum*. (**B**) Distribution of *S. maydis* worldwide. Red triangles represent collection locations reported in the literature. (**C**) *Serratia symbiotica* (red) is compartmentalized into syncytial secondary bacteriocytes (Sba) sandwiched between the uninucleate primary bacteriocytes (Pba) housing *Buchnera aphidicola* (green), forming a horseshoe-shaped bacteriome (green, red, and blue signals indicate *B. aphidicola* cells, *S. symbiotica* cells and host insect nuclei, respectively). (**D**) Close-up view of primary and secondary bacteriocytes showing their embedded layout.

the difficulty of rearing some of these insects (e.g. cicadas), the absence of a clonal phase enabling individuals of identical genotypes to be obtained (e.g. cicadas, psyllids, whiteflies, leafhoppers), and their very small size (e.g. psyllids, whiteflies). A candidate species that overcomes all these difficulties is the cereal aphid *S. maydis* (Chaitophorinae) (Fig. 1A), a species that feeds on many species of grass (Poaceae) and is distributed in Europe, much of Asia and has recently reached North and South America where it is considered an invasive species that can damage cereal crops (Fig. 1B)[42]. *S. maydis* is easy to collect in the field, to rear and to reproduce clonally, making it an ideal species for experimental studies. Another advantage of this aphid species is that the genomes of the ancient obligate symbiont *Buchnera aphidicola* and the more recently acquired co-obligate symbiont *Serratia symbiotica* have recently been sequenced and annotated[37]. The two nutritional symbionts are compartmentalized in distinct bacteriocytes: *S. symbiotica* is confined to large syncytial secondary bacteriocytes sandwiched between uninucleate primary bacteriocytes containing *B. aphidicola* (Fig. 1C-D). This case of dual endosymbiosis is particularly relevant for studying how nutritional symbionts dwelling in distinct but contiguous bacteriocytes can collaborate metabolically with each other and with the host. However, the study of this valuable symbiotic system suffers from a lack of genomic information on the insect host. Assembling and annotating the complete genome sequence of *S. maydis* would be a highly informative and fruitful resource for deciphering multiple aspects of the species' biology, and in particular the processes governing cooperative life between insects and anatomically integrated symbionts that form a metabolic unit in a three-way mutualistic symbiosis.

We present here the first complete assembly of the cereal aphid *S. maydis* using a PacBio high fidelity (HiFi) approach. The final assembly is 409.54 Mb in length, with a scaffold N50 of 187.22 kb and 95.5% completeness, providing an excellent genomic resource for further research on *S. maydis*. Structural annotation reveals that the genome contains 29% repeat sequences, and 24,453 protein-coding genes. Functional annotation focused on metabolism and metabolic pathway reconstruction, identifying the 2,541 enzymes of *S. maydis* involved in 273 metabolic pathways. These genomic and metabolic data provide a unique tool for studying the influence of bacterial symbiosis on insect genome evolution, and for exploring in depth the biology of *S. maydis*, an in particular the mechanisms underpinning an interdependent tripartite collaborative life between an insect and its prokaryotic partners.

Methods

Sample collection and genome sequencing. A colony of *S. maydis* sampled on *Hordeum vulgare* in Midelt (Morocco) in April 2016 was used to generate a clonal line from a single individual. Aphids were reared on *Triticum aestivum* (bread wheat) under long-day conditions (16 h light, 8 h dark) in a room maintained at a constant temperature of 20 °C to ensure parthenogenic reproduction. Thirty adult individuals were used for DNA extraction. Only the heads were used to minimize DNA contamination by the symbionts *B. aphidicola* and *S. symbiotica* that are present only in the abdomen, and whose genomes have been sequenced previously from the same aphid clonal line[37]. Whole insects were first surface sterilized with 99% ethanol, rinsed with sterile water and then immersed in 70% ethanol where the heads were dissected with microscissors before being stored directly in a

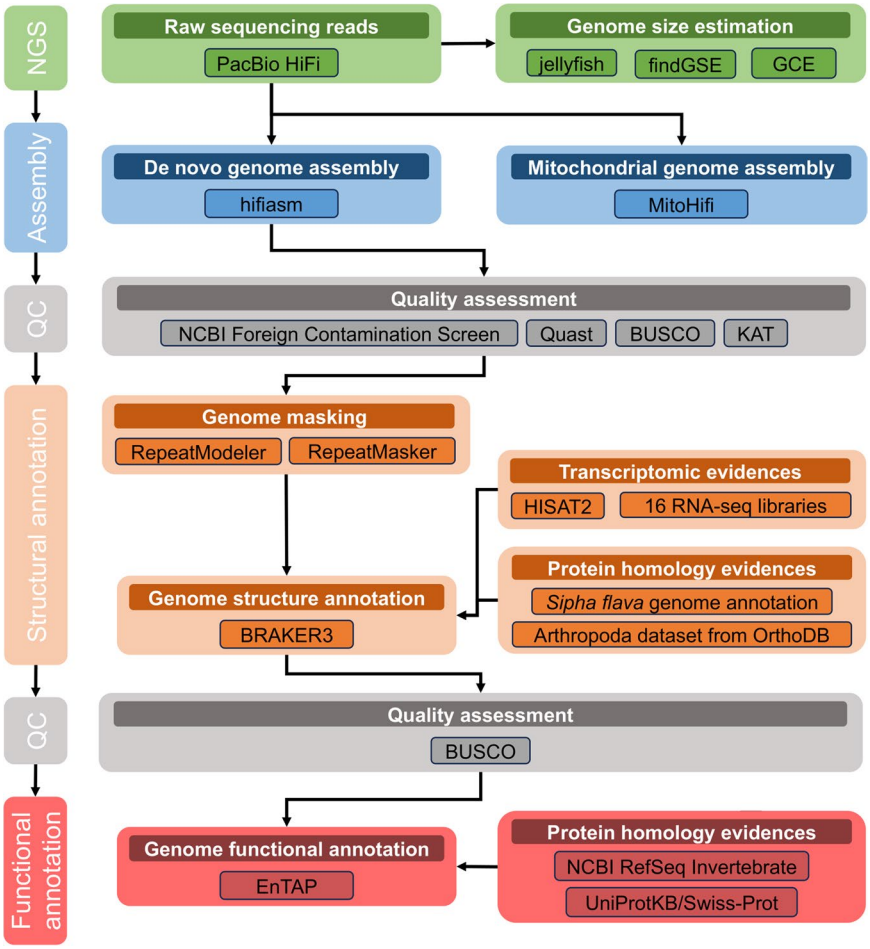

Fig. 2 Flowchart highlighting the *Sipha maydis* genome assembly and annotation process, including quality assessment steps.

sterile plastic tube at −80 °C prior to DNA extraction. DNA extraction was performed using phenol-chloroform. Briefly, tissues were homogenized in 500 ml STE buffer (100 mM NaCl, 1 mM EDTA, 10 mM Tris-Cl, pH 8.0) with a sterile pestle, then treated with 25 µl SDS 10% and 3 µl proteinase K (20 mg/ml). After a two-hour incubation at 55 °C with frequent mixing, the sample was treated with 6 µl RNase (10 mg/ml) and incubated at 37 °C for 30 min. Genomic DNA was purified by two successive extractions with phenol:chloroform:isoamyl alcohol (25:24:1, v/v/v) followed by extraction with 1 vol of chloroform:isoamyl alcohol (24:1, v/v/v). Genomic DNA was then precipitated by 0.7 volumes of isopropanol. After washing the pellet with 70% ethanol, genomic DNA was recovered in TE buffer (1 mM EDTA, 10 mM Tris HCl pH 8). DNA concentrations and quality were assessed using NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA), agarose gel electrophoresis and Qubit fluorometer (Thermo Fisher Scientific). Sequencing libraries were prepared using the Express Template Prep Kit 3.0 (Pacific Biosciences, Menlo Park, USA) and whole-genome sequencing was performed on the PacBio Sequel IIe system at the Genomics Core Leuven (KU Leuven, Leuven, Belgium) using the Sequel® II Binding Kit 3.2 (Pacific Biosciences).

Genome assembly and evaluation. The complete *S. maydis* genome assembly and annotation workflow, including quality assessment steps, is shown in Fig. 2.

The genome size of *S. maydis* was estimated using k-mer analyses from raw PacBio HiFi reads. A k-mer (k = 21) distribution was generated with Jellyfish[43] (v2.2.10) using the PacBio HiFi reads and genome size was estimated using three different strategies: i) findGSE[44] v1.94.R, ii) gce[45] v1.0.2 and iii) the ratio of total distinct k-mers divided by the frequency mode of the k-mer distribution using R[46] v4.2.1 as described in Hon *et al.*[47]. The *S. maydis* genome was assembled from the PacBio HiFi raw reads using hifiasm[48] v0.18.9-r527 with default parameters. The primary assembly was then screened for contaminants using the NCBI Foreign Contamination Screen (FCS) and seven contaminant scaffolds corresponding to the two endosymbiont genomes (*B. aphidicola* and *S. symbiotica*) were removed from the assembly prior to the annotation step. The accuracy and completeness of the assembly were assessed using (i) QUAST[49] v5.0.2 with the –large and -k options, (ii) BUSCO v5.4.6[50] using the Insecta ODB10 database, and (iii) KAT[51] v2.4.2 to compute shared k-mers between PacBio HiFi reads and the assembly. A total of 3.70 Gb of PacBio HiFi reads with a mean read length of 6.89 kb were assembled to generate a 409.54 Mb draft genome assembly consisting of 3,570 contigs with a N50 length of 187.22 kb and a largest contig of 1.25 Mb (Table 1).

Metrics	*Sipha maydis* (this study)	*Sipha flava* (GCF_003268045.1)
Total length (Mb)	409.54	353.18
No. of scaffolds/contigs	3,570	1,923
Scaffold/Contig N50 (kb)	187.22	1,686.65
Scaffold/Contig L50	668	67
GC%	29.71	30.00
No. of protein-coding genes	24,466*	13,575
Mean gene length (kb)	4.98	13.00

Table 1. Genome assembly and annotation statistics of *Sipha maydis* as compared to its close relative *Sipha flava*. * Including the 13 mitochondrial protein-coding genes.

Metrics	*Sipha maydis* (this study)		*Sipha flava* (GCF_003268045.1)	
	Assembly	Predicted gene models	Assembly	NCBI *Sipha flava* Annotation Release 100
Complete BUSCOs (%)	95.5	94.8	94.3	94.7
Complete Single-Copy BUSCOs (%)	91.0	90.0	92.5	92.3
Complete Duplicated BUSCOs (%)	4.5	4.8	1.8	2.4
Fragmented BUSCOs (%)	0.8	1.8	0.4	0.7
Missing BUSCOs (%)	3.7	3.4	5.3	4.6

Table 2. BUSCO assessment of *Sipha maydis* genome assembly compared with its close relative *Sipha flava* (Insecta_odb10 scores, n:1,367).

Type		Numbers	Length (bp)	% of the genome
Retroelements	SINEs	140	35489	0.01
	Penelope	1028	298541	0.07
	LINEs	26058	13202799	3.22
	LTR elements	4983	3103658	0.76
	Total	31181	16341946	3.99
DNA transposons		49671	14819687	3.62
Unclassified		201353	68775905	16.79
Satellites		338	188152	0.05
Simple repeats		290115	12778827	3.12
Low complexity		45565	2225158	0.54

Table 3. Repetitive sequences in the genome of *Sipha maydis*.

The assembly size is comparable to the genome size estimate of ~433 Mb using k-mers (findGSE: 446.20 Mb; gce: 421.93 Mb; total distinct k-mers divided by the frequency mode of the k-mer distribution: 431.32 Mb). The genome assembly was found to have a high level of completeness (95.5%). Of the 1,367 Insecta BUSCOs, 91.0% were complete and single-copy, 4.5% complete and duplicated, 0.8% fragmented and 3.7% were missing (Table 2). The alternative haplotype-resolved assemblies produced by Hifiasm have however a reduced total length (347.06 Mb and 329.15 Mb), N50 (70.77 kb and 71.73 kb) and complete BUSCO scores (82.6% and 81.3%) (available at: https://doi.org/10.57745/6RYSBE[52]).

The mitochondrial genome was assembled using the MitoHiFi pipeline[53] to generate a 16,379 bp genome consisting of 37 genes, including 13 protein-coding genes, 2 rRNAs, and 22 tRNAs, with a GC content of 15.43%.

Gene prediction and general functional annotation. A *de novo* repeat library was generated using RepeatModeler[54] v1.0.11. RepeatMasker[55] v4.1.2 was then used with the *de novo* filtered repeat library to identify and soft-mask repeats in the draft assembly prior to annotation. Ultimately, we identified 117.21 Mb of repetitive sequences, accounting for 28.62% of the assembled genome (Table 3).

After masking the repeat sequences, the structural annotation (i.e., gene prediction) was performed using the BRAKER3[56–68] pipeline v3.0.3 using *ab initio* prediction, homology searching and transcriptome-based approaches to predict protein-coding genes. For transcriptome-based prediction, the pipeline used 16 RNAseq libraries (PRJNA1031833)[69] that were aligned to the soft-masked genome using HISAT2[70] v2.2.1. For the homology-based approaches, annotated proteins from *Sipha flava* genome annotation (GCA_003268045.1) and the Arthropoda protein dataset from OrthoDB[71] v11 were downloaded. The final set of protein-coding genes was retrieved from the Augustus predictions[56,57]. The completeness of the annotated protein set was assessed using BUSCO[50] v5.4.6 and the Insecta ODB10 database. A first general run of functional annotations of predicted proteins was carried out using EnTAP[72] v0.10.8-beta. Comparisons were performed against UniProtKB/

Swiss-Prot[73] and NCBI RefSeq invertebrate annotated proteins (https://ftp.ncbi.nlm.nih.gov/refseq/release/invertebrate/) as reference databases. Hence, we identified and soft-masked 28.62% (117.21 Mb) of the *S. maydis* genome as repeated sequences. After masking those repeated sequences, a total of 24,453 protein-coding genes were predicted using a combination of *ab initio*, homology-based and transcriptome-based approaches. The completeness of the gene prediction revealed that 94.8% of BUSCO genes were successfully detected (90.0% are single-copied and 4.8% are duplicated).

Functional annotation of metabolism. We used several methods to perform a functional annotation of the *S. maydis* enzyme set: (i) the online KAAS – KEGG[74] v2.1 automatic annotation server against both "For gene" and "For eukaryotes" representative sets, (ii) the v2 of the PRIAM[75] tool, (iii) the Blast2GO[76] pipeline v3.5 and (iv) the InterProScan[77] v5.56 pipeline with a local installation for faster data generation. These methods generated information such as EC numbers, KEGG Orthology and Gene Ontology related to the protein sequences. All annotations were collected in a SQL database using CycADS[78] and associated with the genomic information data. Default settings were used for software configurations and the BLAST alignments (prior to the Blast2GO analysis) were performed against the curated UniProtKB/Swiss-Prot[79] protein sequence database.

Metabolic network reconstruction. The final step in assessing the quality of the *S. maydis* genome was the reconstruction of its metabolic network, which validated the functional annotation of this organism's enzyme set. This expert reconstruction step also makes the network directly accessible to the scientific community through the exploration of the dedicated metabolic database we make publicly available, or as an additional dataset that can be uploaded by users in suitable formats (e.g., sbml and Biopax). The enriched gene records containing all annotations were extracted from the CycADS SQL to generate the corresponding BioCyc-like metabolic database using Pathway Tools v27.0[80,81], which we named SipmaCyc, according to current convention. In the summary section for each gene/protein resulting page, the information relative to the annotation results was recorded to allow the researchers to evaluate the confidence for each putative function assigned to a protein. Incomplete EC_ numbers (i.e., classes and subclasses) and EC numbers inferred from a single Blast2GO annotation were excluded from the network, even though the annotation remains accessible for users in the gene description.

Since the functional annotation of metabolic pathways in an insect dependent on nutritional symbionts cannot be done without taking into account their metabolic contributions, we validated the *S. maydis* metabolic annotations by assessing their homogeneity and correct integration with those of its symbiotic partners, *B. aphidicola* and *S. symbiotica*, using the CycADS annotation system[78,82]. This led to the production of an integrated metabolic network of *S. maydis* and its bacterial associates, which we made publicly available on ArtSymbioCyc[82] (http://artsymbiocyc.cycadsys.org/), a collection of metabolic databases dedicated to arthropod symbioses. *S. maydis* encodes 26,059 predicted proteins from its 24,466 protein-coding genes, of which 2,523 are enzymes involved in 273 metabolic pathways. The SipmaCyc database in ArtSymbioCyc provides a complete description of the central metabolism of *S. maydis* at the genome scale (Fig. 3) and enables users to visualize and explore individual metabolic networks at the level of compounds, reactions, or pathways.

Data Records

The sequencing data that were used for the genome assembly and annotation have been deposited in the NCBI Sequence Read Archive with accession numbers SRP443918[83] and SRP468263[69] respectively. This genome assembly is available under accession number GCA_034509805.1[84].

Data on metabolic network reconstructions are available at Recherche Data Gouv (https://doi.org/10.57745/6RYSBE)[52] as well as in the ArtSymbioCyc collection (http://artsymbiocyc.cycadsys.org/)[82]. Metabolic network reconstructions and the resulting BioCyc metabolism databases are available in the ArthropodaCyc collection[82] (https://arthropodacyc.cycadsys.org) for *S. maydis* alone (organism database: "*Sipha maydis*") and in the ArtSymbioCyc collection (http://artsymbiocyc.cycadsys.org/) for *S. maydis* associated with its two obligate nutritional symbionts (organism databases: "*Sma-Sipha maydis*" and "*Sma-Sipha maydis holobiont*"). A single repository (Recherche Data Gouv)[52] was created to unite (i) the *S. maydis* genome assembly commands file (txt), (ii) the genomic files (fasta primary and haplotype-resolved assemblies), (iii) the gtf/gb structural annotations for both the genome and the mitochondrion and (iv) the functional annotations (tabular text), reactions (sbml) and network in Biopax format for the three partners composing the symbiotic system.

Technical Validation

The quality of the *S. maydis* genome assembly was assessed by computing several metrics: (i) comparison with the estimated genome size, which is ~433 Mb (see above); (ii) genomic BUSCO analyses, which identified 95.5% of Insecta BUSCOs in the *S. maydis* genome (91.0% are single-copied and 4.5% are duplicated, Table 2), and 94.8% of Insecta BUSCOs proteins in its predicted gene models (90.0% are single-copied and 4.8% are duplicated, Table 2); (iii) comparison with the PacBio HiFi reads using QUAST and KAT which showed that 99.95% of the k-mers ($k = 27$) of our assembly are covered by the k-mers from the PacBio HiFi reads and 99.73% of the PacBio HiFi reads could be mapped into the assembly. Despite a low sequencing yield (3.70 Gb) and small read lengths (6.89 kb), the aforementioned quality metrics indicated that the *S. maydis* genome assembly has a high level of completeness and is of high-quality.

The functional annotation of the *S. maydis* genome coding for central metabolism is supported by the use of our CycADS expert system and leads to the possibility of reconstructing the metabolic network, whose integrity and consistency can be tested using the comparative tools of the ArtSymbioCyc database interface[82]. As an

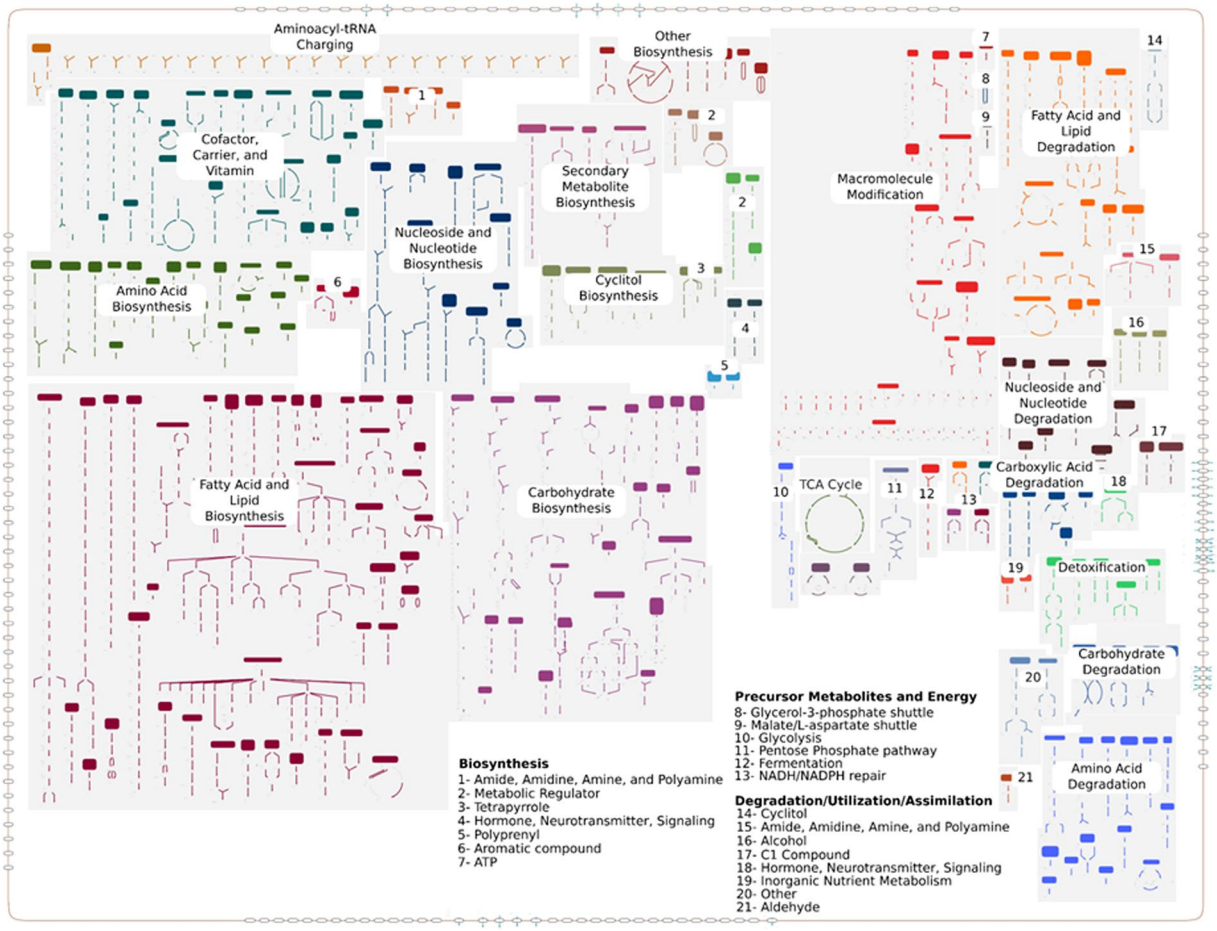

Fig. 3 Schematic overview of the *S. maydis* metabolic network. The figure contains all the 38 metabolic categories, of which only 21 are highlighted for a more reader-friendly representation. Users can explore this map in ArtSymbioCyc, also having access to all reactions and metabolites in each pathway.

EC Category*	*Sipha maydis* (this study)
1–Oxidoreductases	439 (23%)
2–Transferases	709 (38%)
3–Hydrolases	435 (23%)
4–Lyases	133 (7%)
5–Isomerases	69 (4%)
6–Ligases	97 (5%)
7–Translocases	44 (2%)
Total reactions with full or partial EC Numbers	1,882

Table 4. Distribution of *S. maydis* reactions across the 6 top-level categories identified by the Enzyme Commission. *Included in this table are all reactions in the database which have been assigned either full or partial EC numbers, and for which an enzyme has been identified.

example, Table 4 shows the distribution of *S. maydis* reactions in the top 6 levels of the Enzyme Commission classification, which is fully consistent with those of the other insects in the database.

Code availability

All software and pipelines were executed according to the manual and protocols of the published bioinformatic tools. The version and code/parameters of software have been described in the Methods section. Metabolic network reconstructions were carried out using pathway tools 27.0 (April 12, 2023), with annual updates planned.

Received: 17 December 2023; Accepted: 23 April 2024;
Published online: 04 May 2024

References

1. Sudakaran, S., Kost, C. & Kaltenpoth, M. Symbiont acquisition and replacement as a source of ecological innovation. *Trends in Microbiology* **25**, 375–390 (2017).
2. Zientz, E., Dandekar, T. & Gross, R. Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiology and Molecular Biology Reviews* **68**, 745–770 (2004).
3. Baumann, P. Biology of bacteriocyte-associated endosymbionts of plant sap-sucking insects. *Annual Review of Microbiology* **59**, 155–189 (2005).
4. Whittle, M., Barreaux, A. M. G., Bonsall, M. B., Ponton, F. & English, S. Insect-host control of obligate, intracellular symbiont density. *Proceedings of the Royal Society B: Biological Sciences* **288**, 20211993 (2021).
5. Simonet, P. *et al*. Bacteriocyte cell death in the pea aphid/Buchnera symbiotic system. *PNAS* **115**, E1819–E1828 (2018).
6. Bennett, G. M. & Moran, N. A. Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. *PNAS* **112**, 10169–10176 (2015).
7. Douglas, A. E. How multi-partner endosymbioses function. *Nat Rev Microbiol* **14**, 731–743 (2016).
8. McCutcheon, J. P. & von Dohlen, C. D. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Curr Biol* **21**, 1366–1372 (2011).
9. Li, N.-N. *et al*. Bacteriocyte development is sexually differentiated in *Bemisia tabaci*. *Cell Reports* **38**, 110455 (2022).
10. Garber, A. I. *et al*. The evolution of interdependence in a four-way mealybug symbiosis. *Genome Biology and Evolution* **13**, evab123 (2021).
11. Gottlieb, Y. *et al*. Inherited intracellular ecosystem: symbiotic bacteria share bacteriocytes in whiteflies. *The FASEB Journal* **22**, 2591–2599 (2008).
12. Łukasik, P. *et al*. Multiple origins of interdependent endosymbiotic complexes in a genus of cicadas. *PNAS* **115**, E226–E235 (2018).
13. Nakabachi, A. *et al*. Defensive bacteriome symbiont with a drastically reduced genome. *Curr Biol* **23**, 1478–1484 (2013).
14. Manzano-Marín, A., Szabó, G., Simon, J.-C., Horn, M. & Latorre, A. Happens in the best of subfamilies: Establishment and repeated replacements of co-obligate secondary endosymbionts within Lachninae aphids. *Environmental Microbiology* **19**, 393–408 (2017).
15. von Dohlen, C. D. *et al*. Dynamic acquisition and loss of dual-obligate symbionts in the plant-sap-feeding Adelgidae (Hemiptera: Sternorrhyncha: Aphidoidea). *Front Microbiol* **8**, (2017).
16. Kobiałka, M., Michalik, A., Szwedo, J. & Szklarzewicz, T. Diversity of symbiotic microbiota in Deltocephalinae leafhoppers (Insecta, Hemiptera, Cicadellidae). *Arthropod Structure & Development* **47**, 268–278 (2018).
17. Nakabachi, A., Piel, J., Malenovský, I. & Hirose, Y. Comparative genomics underlines multiple roles of *Profftella*, an obligate symbiont of psyllids: Providing toxins, vitamins, and carotenoids. *Genome Biology and Evolution* **12**, 1975–1987 (2020).
18. Sloan, D. B. & Moran, N. A. Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. *Mol Biol Evol* **29**, 3781–3792 (2012).
19. Hall, A. A. G. *et al*. Codivergence of the primary bacterial endosymbiont of psyllids versus host switches and replacement of their secondary bacterial endosymbionts. *Environmental Microbiology* **18**, 2591–2603 (2016).
20. Rao, Q. *et al*. Genome reduction and potential metabolic complementation of the dual endosymbionts in the whitefly *Bemisia tabaci*. *BMC Genomics* **16**, 226 (2015).
21. Santos-Garcia, D. *et al*. To B or Not to B: Comparative genomics suggests *Arsenophonus* as a source of b vitamins in whiteflies. *Front Microbiol* **9**, (2018).
22. Zchori-Fein, E., Lahav, T. & Freilich, S. Variations in the identity and complexity of endosymbiont combinations in whitefly hosts. *Front Microbiol* **5**, (2014).
23. Wang, Y.-B. *et al*. Intracellular symbionts drive sex ratio in the whitefly by facilitating fertilization and provisioning of B vitamins. *ISME J* **14**, 2923–2935 (2020).
24. Husnik, F. & McCutcheon, J. P. Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. *PNAS* **113**, E5416–E5424 (2016).
25. Szabó, G. *et al*. Convergent patterns in the evolution of mealybug symbioses involving different intrabacterial symbionts. *ISME J* **11**, 715–726 (2017).
26. Koga, R., Nikoh, N., Matsuura, Y., Meng, X.-Y. & Fukatsu, T. Mealybugs with distinct endosymbiotic systems living on the same host plant. *FEMS Microbiology Ecology* **83**, 93–100 (2013).
27. Koga, R. & Moran, N. A. Swapping symbionts in spittlebugs: Evolutionary replacement of a reduced genome symbiont. *ISME J* **8**, 1237–1246 (2014).
28. Matsuura, Y. *et al*. Recurrent symbiont recruitment from fungal parasites in cicadas. *PNAS* **115**, E5970–E5979 (2018).
29. Michalik, A. *et al*. Alternative transmission patterns in independently acquired nutritional cosymbionts of Dictyopharidae planthoppers. *mBio* **12**, e01228–21 (2021).
30. Dial, D. T. *et al*. Transitional genomes and nutritional role reversals identified for dual symbionts of adelgids (Aphidoidea: Adelgidae). *ISME J* **16**, 642–654 (2022).
31. Szabó, G., Schulz, F., Manzano-Marín, A., Toenshoff, E. R. & Horn, M. Evolutionarily recent dual obligatory symbiosis among adelgids indicates a transition between fungus- and insect-associated lifestyles. *ISME J* **16**, 247–256 (2022).
32. Toenshoff, E. R., Gruber, D. & Horn, M. Co-evolution and symbiont replacement shaped the symbiosis between adelgids (Hemiptera: Adelgidae) and their bacterial symbionts. *Environmental Microbiology* **14**, 1284–1295 (2012).
33. Weglarz, K. M., Havill, N. P., Burke, G. R. & von Dohlen, C. D. Partnering with a pest: Genomes of hemlock woolly adelgid symbionts reveal atypical nutritional provisioning patterns in dual-obligate bacteria. *Genome Biology and Evolution* **10**, 1607–1621 (2018).
34. von Dohlen, C. D. *et al*. Diversity of proteobacterial endosymbionts in hemlock woolly adelgid (*Adelges tsugae*) (Hemiptera: Adelgidae) from its native and introduced range. *Environmental Microbiology* **15**, 2043–2062 (2013).
35. Toenshoff, E. R., Szabó, G., Gruber, D. & Horn, M. The pine bark adelgid, *Pineus strobi*, contains two novel bacteriocyte-associated gammaproteobacterial symbionts. *Applied and Environmental Microbiology* **80**, 878–885 (2014).
36. Manzano-Marín, A. & Latorre, A. Snapshots of a shrinking partner: Genome reduction in *Serratia symbiotica*. *Scientific Reports* **6**, 32590 (2016).
37. Renoz, F. *et al*. The di-symbiotic systems in the aphids *Sipha maydis* and *Periphyllus lyropictus* provide a contrasting picture of recent co-obligate nutritional endosymbiosis in aphids. *Microorganisms* **10**, 1360 (2022).
38. Renoz, F. *et al*. Compartmentalized into bacteriocytes but highly invasive: The puzzling case of the co-obligate symbiont *Serratia symbiotica* in the aphid *periphyllus lyropictus*. *Microbiol Spectr* e0045722 (2022).
39. Monnin, D. *et al*. Parallel Evolution in the Integration of a Co-obligate Aphid Symbiosis. *Current Biology* **30**, 1949–1957.e6 (2020).
40. Manzano-Marín, A. *et al*. Co-obligate symbioses have repeatedly evolved across aphids, but partner identity and nutritional contributions vary across lineages. *Peer Community Journal* **3**, (2023).
41. Yorimoto, S., Hattori, M., Kondo, M. & Shigenobu, S. Complex host/symbiont integration of a multi-partner symbiotic system in the eusocial aphid *Ceratovacuna japonica*. *iScience* **25**, 105478 (2022).
42. Wieczorek, K. & Bugaj-Nawrocka, A. Invasive aphids of the tribe Siphini: a model of potentially suitable ecological niches. *Agricultural and Forest Entomology* **16**, 434–443 (2014).
43. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

44. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* **34**, 550–557 (2018).
45. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv.org* https://arxiv.org/abs/1308.2012v2 (2013).
46. Team, R. A language and environment for statistical computing. *Computing* **1**, (2006).
47. Hon, T. *et al*. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* **7**, 399 (2020).
48. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
49. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
50. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* **38**, 4647–4654 (2021).
51. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).
52. Renoz, F. *et al*. Genetic and metabolic resources for *Sipha maydis* multi-symbiotic system. *Recherche Data Gouv* https://doi.org/10.57745/6RYSBE (2023).
53. Uliano-Silva, M. *et al*. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics* **24**, 288 (2023).
54. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS* **117**, 9451–9457 (2020).
55. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4, 4.10.1–4.10.14 (2009).
56. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
57. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* **3**, lqaa108 (2021).
58. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-Genome Annotation with BRAKER. *Methods Mol Biol* **1962**, 65–95 (2019).
59. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**, e119 (2014).
60. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform* **2**, lqaa026 (2020).
61. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* **33**, 6494–6506 (2005).
62. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
63. Gotoh, O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res* **36**, 2630–2638 (2008).
64. Iwata, H. & Gotoh, O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res* **40**, e161 (2012).
65. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**, ISCB Comm J-304 (2020).
66. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
67. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
68. Gabriel, L. *et al*. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv* 2023.06.10.544449, https://doi.org/10.1101/2023.06.10.544449 (2023).
69. *NCBI Sequence Read Archive* http://identifiers.org/ncbi/insdc.sra:SRP468263 (2023).
70. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
71. Kuznetsov, D. *et al*. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res* **51**, D445–D451 (2023).
72. Hart, A. J. *et al*. EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol Ecol Resour* **20**, 591–604 (2020).
73. UniProt Consortium UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523–D531 (2023).
74. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35**, W182–185 (2007).
75. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* **31**, 6633–6639 (2003).
76. Conesa, A. & Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* **2008**, 619832 (2008).
77. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
78. Vellozo, A. F. *et al*. CycADS: an annotation database system to ease the development and update of BioCyc databases. *Database (Oxford)* **2011**, bar008 (2011).
79. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. in *Plant Bioinformatics: Methods and Protocols* (ed. Edwards, D.) 89–112, https://doi.org/10.1007/978-1-59745-535-0_4 (Humana Press, Totowa, NJ, 2007).
80. Karp, P. D. *et al*. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform* **22**, 109–126 (2021).
81. Karp, P. D. *et al*. The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* **20**, 1085–1093 (2019).
82. Baa-Puyoulet, P. *et al*. ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods. *Database (Oxford)* **2016**, baw081 (2016).
83. *NCBI Sequence Read Archive* http://identifiers.org/ncbi/insdc.sra:SRP443918 (2023).
84. Renoz, F. *et al*. PacBio Hi-Fi genome assembly of *Sipha maydis*, a model for the study of multipartite mutualism in insects. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_034509805.1 (2023).

Acknowledgements

We thank Karen Gaget and Agnès Vallier (INRAE / INSA Lyon BF2i) for their technical advice on DNA extraction. We also thank Annelien Verfaillie and Wim Meert (Genomics Core Leuven) for their technical advice on the PacBio® HiFi sequencing method. This study was financially supported by FNRS grant no. J.0082.23. This paper is publication BRC 409 of the Biodiversity Research Centre (Université catholique de Louvain). Research at BF2i was supported by the Institut National de la Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE) and the Institut National des Sciences Appliquées de Lyon (INSA Lyon).

Author contributions

F.R., T.H. and F.C. conceived the project. F.R. secured funding for the project. S.F. collected the aphids in Morocco and established the clonal line in the laboratory. F.R. maintained the aphids, collected the samples, extracted the DNA and initiated the sequencing process. N.P., P.B.-P., L.G. and H.C. analyzed the data. All the authors contributed to the drafting of the manuscript. All the authors read, revised and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.R., N.P. or F.C.

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.