Annotated Draft Genome Sequence of the Apple Scab Pathogen *Venturia inaequalis*

Thomas A. J. Passey,[a,b] Andrew D. Armitage,[a] Xiangming Xu[a]

[a]NIAB EMR, East Malling, Kent, United Kingdom
[b]School of Agriculture, Policy and Development, University of Reading, Reading, United Kingdom

ABSTRACT Apple scab is one of the most economically important diseases of apples worldwide. The disease is caused by the haploid ascomycete *Venturia inaequalis*. We present here an annotated *V. inaequalis* whole-genome sequence of 72 Mb, assembled into 238 contigs, with 13,761 predicted genes.

Venturia inaequalis (phylum *Ascomycota*, class *Dothideomycetes*) is the causal agent of apple scab, one of the most important diseases of apples worldwide, and, as a result, has been extensively researched for well over a century (1). If not managed, annual epidemics can result in large numbers of unmarketable fruit. Previously published annotated genome sequences for *V. inaequalis* have between 1,012 and 1,680 scaffolds (2, 3).

A single-spore isolate of *V. inaequalis* (05/172) was obtained in 2005 from a lesion on a leaf of *Malus x domestica* cv. Worcester Pearmain from Ash Farm in Worcestershire, United Kingdom (4). DNA was extracted and sequenced by two methods: (i) DNA was extracted from mycelium using a Qiagen Genomic-tip 100/G kit; the tissue method of sample preparation was used according to the manufacturer's protocol with options 3B and 4B (adapted to 200 μl proteinase K). Isolation of DNA followed the manufacturer's protocol with options 5B and 6B. DNA was sent to the Earlham Institute (Norwich, UK), for sequencing using the Pacific Biosciences (PacBio) platform. (ii) DNA of the isolate was extracted for Passey et al. (5). Paired-end genomic libraries were prepared using a NEXTflex Rapid DNA-Seq version 14.02 library prep kit (Bioo Scientific) following the manufacturer's protocol but modified by using Illumina adapters rather than NEXTflex barcodes. Libraries were validated using a fragment analyzer (Advanced Analytical Technologies), which confirmed a high proportion of library DNA fragments between 600 and 900 bp long. Libraries were sequenced using 2 × 300-bp reads on an Illumina MiSeq platform. Illumina adapters and low-quality base pairs were trimmed from 1,281,750 MiSeq reads with fastq-mcf version 1.04.636 (6).

PacBio sequencing reads (944,907 reads) were corrected, trimmed, and assembled with Canu version 1.2 (7), and the assembly was corrected with MiSeq reads using Pilon version 1.17 (8). Hybrid assembly with both PacBio and MiSeq reads was performed with SPAdes version 3.9.0 (9) and then merged with the Canu assembly using quickmerge version 0.2 (10); the merged assembly was corrected with the MiSeq reads using Pilon. The genome was assembled into 72.3 Mb in 238 contigs (Table 1). Repetitive and low-complexity regions of the merged assembly were identified by repeat masking with RepeatMasker version 4.0.6 (http://www.repeatmasker.org) and TransposonPSI (release 08222010; http://transposonpsi.sourceforge.net), masking 34.2 Mb (47.3%) of the genome, of which 98.7% was due to transposable elements. Quality of the genome assembly was assessed by looking for benchmarking universal single-copy orthologs (BUSCO) with BUSCO version 3 (11) against the *Ascomycota* odb9 data set, identifying 1,286 (out of 1,315) as present in the assembly. Gene prediction was performed with

TABLE 1 *V. inaequalis* isolate 05/172 genome assembly statistics[a]

Statistic	Value for isolate 05/172
No. of contigs	238
Total length (bp)	72,310,420
Largest contig (bp)	3,847,617
GC content (%)	42.75
N_{50} (bp)	953,805
N_{75} (bp)	531,805
L_{50}	23
L_{75}	49

[a]Assembly produced by merged Canu and SPAdes assemblies using PacBio- and MiSeq-generated sequencing reads.

the use of RNA sequencing (RNA-seq) data from Thakur et al. (12); RNA-seq data were aligned to the genome by STAR version 2.6 (13). A predicted 13,761 genes are present in the assembled genome; 11,597 genes were predicted by Braker1 (14), supplemented by 2,164 genes predicted by CodingQuarry (15) (in pathogen mode) in the intergenic regions of Braker1 gene models. Functional annotation of the genome was performed using Interproscan version 5.18-57.0 (16) and the July 2016 release of the Swiss-Prot database (17).

Data availability. The Sequence Read Archive accession numbers are SRR5183052 for the Illumina MiSeq reads and SRR5183051 for the PacBio reads. This whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number QFBF00000000 (BioProject number PRJNA354841). The version described in this paper is the first version, QFBF01000000.

REFERENCES

1. MacHardy WE. 1996. Apple scab: biology, epidemiology, and management. American Phytopathological Society, St Paul, MN.
2. Shiller J, Van de Wouw AP, Taranto AP, Bowen JK, Dubois D, Robinson A, Deng CH, Plummer KM. 2015. A large family of *AvrLm6*-like genes in the apple and pear scab pathogens, *Venturia inaequalis* and *Venturia pirina*. Front Plant Sci 6:980. https://doi.org/10.3389/fpls.2015.00980.
3. Deng CH, Plummer KM, Jones DAB, Mesarich CH, Shiller J, Taranto AP, Robinson AJ, Kastner P, Hall NE, Templeton MD, Bowen JK. 2017. Comparative analysis of the predicted secretomes of Rosaceae scab pathogens *Venturia inaequalis* and *V. pirina* reveals expanded effector families and putative determinants of host range. BMC Genomics 18:339. https://doi.org/10.1186/s12864-017-3699-1.
4. Xu X, Harvey N, Roberts A, Barbara D. 2013. Population variation of apple scab (*Venturia inaequalis*) within mixed orchards in the UK. Eur J Plant Pathol 135:97–104. https://doi.org/10.1007/s10658-012-0068-4.
5. Passey TAJ, Shaw MW, Xu X-M. 2016. Differentiation in populations of the apple scab fungus *Venturia inaequalis* on cultivars in a mixed orchard remain over time. Plant Pathol 65:1133–1141. https://doi.org/10.1111/ppa.12492.
6. Aronesty E. 2013. Comparison of sequencing utility programs. Open Bioinforma J 7:1–8. https://doi.org/10.2174/1875036201307010001.
7. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116.
8. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal.pone.0112963.
9. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
10. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res 44:e147. https://doi.org/10.1093/nar/gkw654.
11. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
12. Thakur K, Chawla V, Bhatti S, Swarnkar MK, Kaur J, Shankar R, Jha G. 2013. *De novo* transcriptome sequencing and analysis for *Venturia inaequalis*, the devastating apple scab pathogen. PLoS One 8:e53937. https://doi.org/10.1371/journal.pone.0053937.
13. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635.
14. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32:767–769. https://doi.org/10.1093/bioinformatics/btv661.

15. Testa AC, Hane JK, Ellwood SR, Oliver RP. 2015. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics 16:170. https://doi.org/10.1186/s12864-015-1344-4.

16. Jones P, Binns D, Chang H-Y, Fraser M, Li W, Mcanulla C, Mcwilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

17. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. 2016. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view, p. 23–54. In Plant bioinformatics: methods and protocols, methods in molecular biology, vol 1374. Humana Press, New York, NY.