Genome Update of *Botrytis cinerea* Strains B05.10 and T4

Martijn Staats[a] and Jan A. L. van Kan[b]

Biosystematics Group, Wageningen University, Wageningen, the Netherlands,[a] and Laboratory of Phytopathology, Wageningen University, Wageningen, the Netherlands[b]

We report here an update of the *Botrytis cinerea* strains B05.10 and T4 genomes, as well as an automated preliminary gene structure annotation. High-coverage *de novo* assemblies and reference-based alignments led to a correction of wrong base calls, elimination of sequence gaps, and improved joining of contigs. The new assemblies have substantially lower numbers of scaffolds and a concomitant increase in the N_{50}.The list of protein-coding genes was generated using the evidence-driven gene predictor Augustus, with expressed sequence tag evidence and RNA-Seq data as input.

The fungal pathogen *Botrytis cinerea* (teleomorph *Botryotinia fuckeliana*) causes serious losses in more than 200 crop species worldwide (2, 14). The development of strategies for its control are difficult, due to the broad genetic plasticity of the fungus (14). *B. cinerea* genome sequences have proven to be essential in dissecting the genetic basis of pathogenicity (4, 11). The first genome assemblies of B05.10 and T4 were sequenced using Sanger technology at low coverage (1). The draft assembly of B05.10 consisted of 588 scaffolds (N_{50}, 257 kb), and ~8% of the assembly was comprised of interscaffold gaps. The draft assembly of T4 consisted of 118 scaffolds, with an N_{50} of 562 kb. The sequence gaps and the large number of (potentially) spurious gene models in the first assemblies of the B05.10 and T4 genomes (1) seriously hampered comparative genome analyses. Therefore, our aim was to sequence the B05.10 and T4 genomes *de novo* and to integrate scaffolds into superscaffolds by using reference-based alignments.

B. fuckeliana strains B05.10 and T4 were cultivated on microporous membranes (pore diameter, 2.4 nm) overlaying malt extract medium. The mycelium was harvested, lyophilized, and ground into a powder. Genomic DNA was extracted using a modified cetyltrimethylammonium bromide method (3) and treated with RNase (Qiagen). Paired-end 100-cycle multiplex sequencing was performed using the Illumina HiSeq2000 technology. For T4, libraries with average insert sizes of 350 bp and 3.5 kb were sequenced by Macrogen Inc. For B05.10, a 150-bp paired-end library (DNAVision) and a 3.5-kb mate pair library (Macrogen) were sequenced. The raw sequence reads were end trimmed to a minimum first-quartile quality score of 28 by using the FASTX tool kit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). The quality-trimmed data consisted of 81.1 million sequencing reads (3.94 Gb) for B05.10 and 114.6 million sequencing reads (6.38 Gb) for T4.

The following procedure was used to build assembly version 2 of B05.10. First, Velvet (15) was run with a hash length 29 to generate the primary *de novo* assembly, using the Illumina reads. Second, superscaffolds were built by manually merging the set of *de novo* scaffolds with the scaffolds derived from the reference assembly (1). Regions with 99% minimum alignment identity between both assemblies were identified using MUMmer (8) and aligned using Geneious (Biomatters Ltd.). In general, the reference assembly was used to extend or connect the scaffolds generated by Velvet and, where possible, to fill in regions containing unresolved characters. Ambiguously aligned regions or regions containing gaps and unresolved characters (i.e., a stretch of Ns in the underlying sequence data) were manually curated. Scaffolds that were chimeric were identified by mapping the Illumina reads back to the superscaffolds by using Bowtie (9) and then manually split. Third, a consensus sequence was generated from the mapped read alignment data by using Geneious. Scaffolds smaller than 25 kb were not included in the final genome assembly of B05.10 (or of T4), as these had substantial assembly problems with many nucleotide gaps. The final genome assembly of B05.10 consists of 82 superscaffolds (N_{50}, 970 kb) with a total size of ~41.2 Mb. The overall G+C content is 42.75%, and there are 0.427 million IUB characters. The same procedure was used to align the Velvet assembly of T4 with the T4 reference genome (1), except that Velvet was run with a hash length 33. The final genome assembly of T4 consists of 56 superscaffolds (N_{50}, 1.71 Mb) with a total size of ~41.6 Mb. The overall G+C content is 42.44%, and there are 0.277 million IUB characters.

We identified and masked repetitive sequences by using RepeatMasker (http://www.repeatmasker.org) and the Repbase fungal repeat database (5). Approximately 1.3% of the B05.10 and T4 genome was found to be composed of repetitive sequences, and the remainder was used for gene finding using Augustus (12). Experimental evidence for gene structures, as provided by 8,954 T4 expressed sequence tag contigs (1) and 1,212 manually curated T4 proteins, were used to generate hints using BLAT (7) and Scipio (6). Additional gene hints were generated from 92.1 million 100-bp Illumina reads for poly(A)-selected RNA from B05.10 grown on glucose, polygalacturonic acid, and tomato leaves (16 and 40 h postinfection) and mixed apothecia and sclerotia of *B. cinerea* strains SAS56 and SAS405. The RNA-Seq reads were mapped with TopHat (13), and transcripts were assembled using CUFFLINKS (10). For B05.10, 1,255 transcripts with an average read depth coverage higher than 200 were used as hints. For T4, 1,306 transcripts were used. Minor manual editing of gene hints was performed to repair any obviously errant gene structures prior to the search for protein-coding genes. For B05.10, the gene-finding strategy resulted in 10,427 protein-coding gene structures, of which 7 (0.07%) included alternatively spliced isoforms and 76

Received 18 August 2012 **Accepted** 23 August 2012

Address correspondence to Jan A. L. van Kan, jan.vankan@wur.nl.

Copyright © 2012, American Society for Microbiology. All Rights Reserved.

doi:10.1128/EC.00164-12

gene models contained internal stop codons. For T4, 10,467 protein-coding genes were predicted, of which 10 (0.10%) included alternatively spliced isoforms and 66 were spurious genes. Spurious gene calls were considered to be nonreliable and were therefore removed from further analyses.

Cuffcompare (10) was used to compare gene structure predictions for B05.10 to those in T4. Overall, 88.6% of the predicted genes (9,211 of 10,401) in T4 had a complete match in B05.10. We found 130 novel genes for B05.10 and 315 novel genes for T4, compared to one another. Using DNAdiff of the MUMmer package, 96.5% of the T4 genome and 97.3% of the B05.10 genome could be aligned. A total of 131,066 insertion/deletion positions and 187,168 single-nucleotide polymorphisms were identified, indicating that the overall rate of difference was similar to that previously reported (1). The draft assembly version 2 of the B05.10 and T4 genomes and gene structure predictions will enable better synteny and orthology analyses and provide a new template for manual curation that will allow the scientific community to eventually close in on a final set of gene annotations and genomic structures for the *B. cinerea* genome.

Nucleotide sequence accession numbers. The results of these whole-genome shotgun projects have been deposited in DDBJ/EMBL/GenBank under accession numbers AAID00000000 and ALOC00000000. The version of B05.10 described in this paper is the second version, AAID02000000. The version of T4 described in this paper is the first version, ALOC01000000. The B05.10 and T4 sequences and predicted gene sets have also been deposited with the Broad Institute *Botrytis cinerea* database, at http://www.broadinstitute.org/annotation/genome/botrytis_cinerea/MultiHome.html.

ACKNOWLEDGMENTS

We thank Cizar Almalak and Nick Alberts for help and contributions.

This research was funded by Wageningen University.

REFERENCES

1. **Amselem J, et al.** 2011. Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. PLoS Genet. **7**:e1002230. doi:10.1371/journal.pgen.1002230.
2. **Dean R, et al.** 2012. The top 10 fungal pathogens in molecular plant pathology. Mol. Plant Pathol. **13**:414–430.
3. **Doyle JJ, Doyle JL.** 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. **19**:11–15.
4. **Gamboa Melendez H, Billon-Grand G, Fevre M, Mey G.** 2009. Role of the *Botrytis cinerea* FKBP12 ortholog in pathogenic development and in sulfur regulation. Fungal Genet. Biol. **46**:308–320.
5. **Jurka J, et al.** 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110**:462–467.
6. **Keller O, Odronitz M, Stanke M, Kollmar M, Waack S.** 2008. Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics **9**:278. doi:10.1186/1471-2105-9-278.
7. **Kent WJ.** 2002. BLAT: the BLAST-like alignment tool. Genome Res. **12**: 656–664.
8. **Kurtz S, et al.** 2004. Versatile and open software for comparing large genomes. Genome Biol. **5**:R12. doi:10.1186/gb-2004-5-2-r12.
9. **Langmead B, Trapnell C, Pop M, Salzberg SL.** 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**:R25. doi:10.1186/gb-2009-10-3-r25.
10. **Roberts A, Pimentel H, Trapnell C, Pachter L.** 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics **27**:2325–2329. doi:10.1093/bioinformatics/btr355.
11. **Schumacher J, Viaud M, Simon A, Tudzynski B.** 2008. The Gα subunit BCG1, the phospholipase C (BcPLC1) and the calcineurin phosphatase co-ordinately regulate gene expression in the grey mould fungus Botrytis cinerea. Mol. Microbiol. **67**:1027–1050.
12. **Stanke M, Diekhans M, Baertsch R, Haussler D.** 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics **24**:637–644.
13. **Trapnell C, Pachter L, Salzberg SL.** 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**:1105–1111.
14. **Williamson B, Tudzynski B, Tudzynski P, van Kan JAL.** 2007. *Botrytis cinerea*: the causes of grey mould disease. Mol. Plant Pathol. **8**:561–580.
15. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.