Draft Genome Sequence and Gene Annotation of the Entomopathogenic Fungus *Verticillium hemipterigenum*

Fabian Horn,[a] Andreas Habel,[b] Daniel H. Scharf,[c] Jan Dworschak,[b] Axel A. Brakhage,[c] Reinhard Guthke,[a] Christian Hertweck,[b] Jörg Linde[a]

Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Jena, Germany[a]; Biomolecular Chemistry, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Jena, Germany[b]; Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Jena, Germany[c]

Verticillium hemipterigenum (anamorph *Torrubiella hemipterigena*) is an entomopathogenic fungus and produces a broad range of secondary metabolites. Here, we present the draft genome sequence of the fungus, including gene structure and functional annotation. Genes were predicted incorporating RNA-Seq data and functionally annotated to provide the basis for further genome studies.

The filamentous fungus *Verticillium hemipterigenum* (anamorph *Torrubiella hemipterigena*) belongs to the phylum *Ascomycota*. *Verticillium* spp. are commonly known as phytopathogenic species (1). *V. hemipterigenum*, however, is an insect pathogen and mainly infects leafhoppers belonging to the family *Cicadellidae* (2). *V. hemipterigenum* strain BCC 1449 produces a number of highly active secondary metabolites like diketopiperazines and enniatins (3–6).

V. hemipterigenum strain BCC 1449 was grown on Czapek-Dox media and mycelium was harvested after 48 h. Three libraries (paired end [PE], 5 kb mate pair, and 8 kb mate pair) were prepared and sequenced using Illumina HiSeq2000 by LGC Genomics (Berlin). Raw sequence reads were quality trimmed, error corrected (7), digitally normalized (8), and assembled with Allpaths-LG (9). Gaps in assemblies were closed using SOAP Gap-Closer (10). For transcriptome sequencing, the fungus was grown under three conditions: Czapek-Dox broth, potato dextrose broth, and Sabouraud dextrose broth from BD (Heidelberg). Transcriptome sequencing was performed at LGC Genomics using HiSeq2000 100 bp PE.

Structural and functional gene annotation was performed as described previously. (11, 12). In short, transcriptome assemblies were generated using Cufflinks (13) and Trinity (14) and mapped back to the reference genome using PASA (15). Parameter sets for *ab initio* gene prediction (Augustus [16] and SNAP [17]) were trained using gene models that were predicted by TransDecoder (15) from aligned transcripts. GeneMark-ES (18) was used without training. Transcriptome data were incorporated into Augustus, FGENESH (19), and PASA. In order to create protein alignments, protein sequences from *V. dahliae* (BROAD) and *V. alfalfae* (BROAD) were mapped. EVidenceModeler (20) was used to combine gene predictions, while PASA (15) was used to predict untranslated regions. Gene functional annotation was performed using Blast2GO (21) and InterproScan (22). SMURF (23) was used to predict secondary metabolite gene clusters. Gene names and functional descriptions were obtained by blasting against the fungal UniProt Knowledgebase (24).

The genome assembly is based on sequencing 7.0 Gbp, which represents an estimated 240-fold genome coverage. The assembly consists of 26 scaffolds with a total size of 28.5 Mbp (N_{50} = 6,006 kbp). Using CEGMA (25), we identified 478 core proteins within the genome. During genome annotation, we utilized RNA-seq data amounting to 21.0 Gbp, which represents an estimated 740-fold genome coverage. Our gene structural annotation pipeline predicted 10,773 genes. Functional annotation resulted in gene ontology categories for 5,845 genes, which have been made available for enrichment analysis with FungiFun2 (26). Additionally, we predicted InterProDomains for 2,749 genes. Prediction of secondary metabolite gene clusters revealed that 404 genes are part of 27 secondary metabolite biosynthesis gene clusters, including 13 PKSs and 16 NRPSs and 4 hybrid PKS-NRPS gene clusters.

Nucleotide sequence accession numbers. This genome project was uploaded to DDBJ/ENA/GenBank and is available under the accession numbers CDHN01000001 to CDHN01000026. This paper describes the first version of the genome. Genome data and additional information are also available at the HKI Genome Resource (http://www.genome-resource.de).

REFERENCES

1. Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BPHJ, Chen Z, Henrissat B, Lee Y-H, Park J, Garcia-Pedrajas MD, Barbara DJ, Anchieta A, de Jonge R, Santhanam P, Maruthachalam K, Atallah Z, Amyotte SG, Paz Z, Inderbitzin P, Hayes RJ, Heiman DI, Young S, Zeng Q, Engels R, Galagan J, Cuomo CA, Dobinson KF, Ma

L-J. 2011. Comparative genomics yields insights into niche adaptation of plant vascular Wilt pathogens. PLoS Pathog **7**:e1002137. http://dx.doi.org/10.1371/journal.ppat.1002137.

2. **Hywel-Jones NL, Evans HC, Jun Y.** 1997. A re-evaluation of the leafhopper pathogen *Torrubiella hemipterigena*, its anamorph *Verticillium hemipterigenum* and *V. pseudohemipterigenum* sp. nov. Mycol Res **101**:1242–1246. http://dx.doi.org/10.1017/S0953756297003912.

3. **Nilanonta C, Isaka M, Chanphen R, Thong-orn N, Tanticharoen M, Thebtaranonth Y.** 2003. Unusual enniatins produced by the insect pathogenic fungus *Verticillium hemipterigenum*: isolation and studies on precursor-directed biosynthesis. Tetrahedron **59**:1015–1020. http://dx.doi.org/10.1016/S0040-4020(02)01631-9.

4. **Nilanonta C, Isaka M, Kittakoop P, Saenboonrueng J, Rukachaisirikul V, Kongsaeree P, Thebtaranonth Y.** 2003. New diketopiperazines from the entomopathogenic fungus *Verticillium hemipterigenum* BCC 1449. J Antibiot **56**:647–651. http://dx.doi.org/10.7164/antibiotics.56.647.

5. **Supothina S, Isaka M, Kirtikara K, Tanticharoen M, Thebtaranonth Y.** 2004. Enniatin production by the entomopathogenic fungus *Verticillium hemipterigenum* BCC 1449. J Antibiot **57**:732–738. http://dx.doi.org/10.7164/antibiotics.57.732.

6. **Isaka M, Palasarn S, Rachtawee P, Vimuttipong S, Kongsaeree P.** 2005. Unique diketopiperazine dimers from the insect pathogenic fungus *Verticillium hemipterigenum* BCC 1449. Org Lett **7**:2257–2260. http://dx.doi.org/10.1021/ol0507266.

7. **Liu Y, Schröder J, Schmidt B.** 2013. Musket: a multistage *k*-mer spectrum-based error corrector for Illumina sequence data. BioInformatics **29**:308–315. http://dx.doi.org/10.1093/bioinformatics/bts690.

8. **Crusoe MR, Edvenson G, Fish J, Howe A, McDonald E, Nahum J, Nanlohy K, Ortiz-Zuazaga H, Pell J, Simpson J, Scott C, Srinivasan RR, Zhang Q, Brown CT.** 2014. The khmer software package: enabling efficient sequence analysis. figshare. http://dx.doi.org/10.6084/m9.figshare.979190.

9. **Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB.** 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A **108**:1513–1518. http://dx.doi.org/10.1073/pnas.1017351108.

10. **Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J.** 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience **1**:18. http://dx.doi.org/10.1186/2047-217X-1-18.

11. **Linde J, Schwartze V, Binder U, Lass-Flörl C, Voigt K, Horn F.** 2014. *De novo* whole-genome sequence and genome annotation of *Lichtheimia ramosa*. Genome Announc **2**(5):e00888-14. http://dx.doi.org/10.1128/genomeA.00888-14.

12. **Linde J, Duggan S, Weber M, Horn F, Sieber P, Hellwig D, Riege K, Marz M, Martin R, Guthke R, Kurzai O.** Defining the transcriptomic landscape of *Candida glabrata* by RNA-Seq. Nucleic Acids Res [Epub ahead of print.] http://dx.doi.org/10.1093/nar/gku1357.

13. **Garber M, Grabherr MG, Guttman M, Trapnell C.** 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods **8**:469–477. http://dx.doi.org/10.1038/nmeth.1613.

14. **Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L.** 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol **28**:511–515. http://dx.doi.org/10.1038/nbt.1621.

15. **Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O.** 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res **31**:5654–5666. http://dx.doi.org/10.1093/nar/gkg770.

16. **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. http://dx.doi.org/10.1093/bioinformatics/btn013.

17. **Korf I.** 2004. Gene finding in novel genomes. BMC Bioinformatics **5**:59. http://dx.doi.org/10.1186/1471-2105-5-59.

18. **Borodovsky M, Lomsadze A.** 2011. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinformat **35**:. http://dx.doi.org/10.1002/0471250953.bi0406s35.

19. **Salamov AA, Solovyev VV.** 2000. Ab initio gene finding in *Drosophila* genomic DNA. Genome Res **10**:516–522. http://dx.doi.org/10.1101/gr.10.4.516.

20. **Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR.** 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol **9**:R7. http://dx.doi.org/10.1186/gb-2008-9-1-r7.

21. **Conesa A, Götz S, Garcia-Gómez JM, Terol J, Talon M, Robles M.** 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21**:3674–3676. http://dx.doi.org/10.1093/bioinformatics/bti610.

22. **Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R.** 2005. InterProScan: protein domains identifier. Nucleic Acids Res **33**:W116–W120. http://dx.doi.org/10.1093/nar/gki442.

23. **Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS, O'Donovan C, Redaschi N, Yeh LS.** 2004. UniProt: the universal protein knowledgebase. Nucleic Acids Res **32**:D115–D119. http://dx.doi.org/10.1093/nar/gkh131.

24. **Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, Fedorova ND.** 2010. SMURF: genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol **47**:736–741. http://dx.doi.org/10.1016/j.fgb.2010.06.003.

25. **Parra G, Bradnam K, Korf I.** 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. BioInformatics **23**:1061–1067. http://dx.doi.org/10.1093/bioinformatics/btm071.

26. **Priebe S, Kreisel C, Horn F, Guthke R, Linde J.** 2014. FungiFun2: a comprehensive online resource for systematic analysis of gene lists from fungal species. Bioinformatics [Epub ahead of print.] http://dx.doi.org/10.1093/bioinformatics/btu627.