RESEARCH ARTICLE

Single-Molecule Real-Time Sequencing Combined with Optical Mapping Yields Completely Finished Fungal Genome

Luigi Faino,[a] Michael F. Seidl,[a] Erwin Datema,[b] Grardy C. M. van den Berg,[a] Antoine Janssen,[b] Alexander H. J. Wittenberg,[b] Bart P. H. J. Thomma[a]

Laboratory of Phytopathology, Wageningen University, Wageningen, The Netherlands[a]; KeyGene N.V., Wageningen, The Netherlands[b]

L.F. and M.F.S. contributed equally to this work.

ABSTRACT **Next-generation sequencing (NGS) technologies have increased the scalability, speed, and resolution of genomic sequencing and, thus, have revolutionized genomic studies. However, eukaryotic genome sequencing initiatives typically yield considerably fragmented genome assemblies. Here, we assessed various state-of-the-art sequencing and assembly strategies in order to produce a contiguous and complete eukaryotic genome assembly, focusing on the filamentous fungus *Verticillium dahliae*. Compared with Illumina-based assemblies of the *V. dahliae* genome, hybrid assemblies that also include PacBio-generated long reads establish superior contiguity. Intriguingly, provided that sufficient sequence depth is reached, assemblies solely based on PacBio reads outperform hybrid assemblies and even result in fully assembled chromosomes. Furthermore, the addition of optical map data allowed us to produce a gapless and complete *V. dahliae* genome assembly of the expected eight chromosomes from telomere to telomere. Consequently, we can now study genomic regions that were previously not assembled or poorly assembled, including regions that are populated by repetitive sequences, such as transposons, allowing us to fully appreciate an organism's biological complexity. Our data show that a combination of PacBio-generated long reads and optical mapping can be used to generate complete and gapless assemblies of fungal genomes.**

IMPORTANCE **Studying whole-genome sequences has become an important aspect of biological research. The advent of next-generation sequencing (NGS) technologies has nowadays brought genomic science within reach of most research laboratories, including those that study nonmodel organisms. However, most genome sequencing initiatives typically yield (highly) fragmented genome assemblies. Nevertheless, considerable relevant information related to genome structure and evolution is likely hidden in those nonassembled regions. Here, we investigated a diverse set of strategies to obtain gapless genome assemblies, using the genome of a typical ascomycete fungus as the template. Eventually, we were able to show that a combination of PacBio-generated long reads and optical mapping yields a gapless telomere-to-telomere genome assembly, allowing in-depth genome analyses to facilitate functional studies into an organism's biology.**

Received 4 June 2015 **Accepted** 24 July 2015 **Published** 18 August 2015

Citation Faino L, Seidl MF, Datema E, van den Berg GCM, Janssen A, Wittenberg AHJ, Thomma BPHJ. 2015. Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. mBio 6(4):e00936-15. doi:10.1128/mBio.00936-15.

Invited Editor Joerg Kaemper, University of Karlsruhe **Editor** Regine Kahmann, MPI for Terrestrial Microbiology

Address correspondence to Bart P. H. J. Thomma, bart.thomma@wur.nl.

Over the last years, the emergence and rapid evolution of whole-genome sequencing technologies have profoundly affected genomic research (1). In the past, whole-genome sequencing projects involved cost-intensive and laborious Sanger sequencing that typically delivered highly fragmented genome assemblies. The high demand for resequencing projects, as well as for *de novo* genome assemblies, has incited the emergence of novel technologies, caught under the umbrella term next-generation sequencing (NGS), which routinely produce gigabases of data in a short amount of time. NGS technologies can be divided in those that generate large amounts of short DNA sequence reads (<500 bp) that are typically characterized by their high quality (1 sequence error per 1 kb) and technologies that produce long reads (>1 kb), although often with relatively low quality (1 to 2 sequence errors per 100 bases) (2). The added value of short reads for *de novo* genome assemblies is typically hampered if repetitive sequences in a genome are longer than the reads themselves, because these reads will be collapsed into a single element, leading to continuity breaks of the assembly at each genomic location that contains the repetitive element (3). Paired reads, consisting of two reads generated from a single DNA fragment and separated by a known distance, can help to increase the continuity of the assembly, provided that the distance between the pair of reads is longer than the repeat itself (3). However, the sequence spanned by these paired reads will often remain unknown, and therefore, *de novo*-produced short-read assemblies will mainly comprise nonrepetitive regions (4, 5). Different sequencing strategies, such as single-molecule real-time (SMRT) and nanopore sequencing, can be used to characterize repeat-rich genomic regions (6–9). Both sequencing strategies generate long reads (up to ~50 kb) that can read through entire repeats and, thus, facilitate more contiguous genome assemblies (6, 10). SMRT sequencing is a well established

TABLE 1 Statistics for the various genome assemblies of *Verticillium dahliae* strain JR2

Source of data, metric	Value for indicated assembly and data sources							
	VerdaJR2v1.5	VerdaJR2v1.5	VerdaJR2v1.5	SPAdes 3.0	SPAdes 3.0	A5	A5	VDAG_JR2v4.0
PE library	X	X	X	X	X	X	X	
MP library	X	X	X	X	X	X	X	
PacBio P4-C2		X		X		X		X
PacBio P5-C3			X		X		X	X
Optical map	X	X	X					X
Contig metrics								
No. of contigs (≥0 bp)	4,514	515	533	2,335	2,463	1,013	1,195	8
No. of contigs (≥1,000 bp)	3,262	324	338	1,579	1,570	415	419	8
Longest contig (bp)	99,830	2,178,335	2,251,806	227,026	543,223	2,308,962	2,304,878	9,275,483
Total length (bp)	33,523,879	35,178,480	35,520,228	34,886,730	35,110,786	36,248,419	36,213,197	36,150,287
N_{50} (bp)	17,466	662,062	649,303	46,943	50,038	598,861	512,741	4,168,633
No. of Ns/100 kb[a]	0	0	0	0	0	0	0	0
Scaffold metrics								
No. of scaffolds (≥0 bp)	9	9	9	1,334	1,510	606	599	8
No. of scaffolds (≥1,000 bp)	9	9	9	659	702	298	285	8
Longest scaffold	9,141,183	9,180,926	9,215,033	1,263,620	1,066,798	2,912,494	2,937,429	9,275,483
Total length	37,537,096	38,353,192	38,703,526	34,780,691	34,969,668	36,425,691	36,548,884	36,150,287
N_{50} (bp)	4,064,734	4,091,407	4,087,047	350,075	306,662	781,486	808,031	4,168,633
No. of Ns/100 kb	10,691.34	8,277.57	8,224.83	109.82	100.64	652.57	1,082.47	0

[a] Ns, unknown nucleotides.

technology that produces long reads (on average, >15 kb and up to ~50 kb) that can be used to improve previously generated genome assemblies using a limited amount of data (~10 to 20× genome coverage) (11) or for *de novo* genome assemblies (≥50× genome coverage) (12). While prokaryote genomes can be completely assembled solely based on long reads (12), the assembly of complete chromosomes of eukaryotic genomes is less straightforward.

Optical mapping is a technique for constructing ordered high-resolution restriction maps from single DNA molecules in a genomewide fashion. As the technique can be used to align *in silico*-generated restriction maps of genome assemblies, optical mapping can be used to direct the placement of individually assembled sequence contigs onto chromosomes (13–15). Nevertheless, optical mapping is not routinely used in genome sequencing initiatives.

Genomic studies have revealed the importance of noncoding regions, structural rearrangements, and repetitive elements for the lifestyle of many organisms (16–18). As these regions are notoriously difficult to assemble in genome sequencing projects, several approaches have been developed to reconstruct such regions (8, 19–21). In this study, we set out to find an optimal strategy to obtain gapless eukaryotic genome assemblies. To this end, we focused on the genome of a filamentous fungal plant pathogen, *Verticillium dahliae*, with a predicted genome size of 36 Mb, as a model. Previous analyses have revealed extensive genomic rearrangements between individual strains of this species and, furthermore, identified distinct genomic regions that are enriched in repetitive elements (18, 22–24). These findings necessitate the improvement of the current fragmented genome assemblies for this species and, thus, provides an excellent target for our study.

RESULTS

Long reads increase genome assembly quality. We previously generated a genome assembly of *V. dahliae* strain JR2 based on paired-end (PE) and mate-pair (MP) reads generated upon Illumina sequencing of short (500-bp) and long (5-kb) insert size libraries (24). In this so-called VerdaJR2v1.5 assembly, optical mapping was used to connect scaffolds, leading to about 4,500 contigs with a contig N_{50} of ~17 kb (Table 1) (22). In order to reduce the number and length of gaps in this assembly, here we used long reads generated with PacBio sequencing for gap filling and scaffolding (see Table S1 in the supplemental material). To this end, two data sets were generated using a conventional (P4-C2 [P4 polymerase and C2 chemistry]) and an improved (P5-C3) version of PacBio chemistry (see Table S1). Sequencing of four SMRT cells with P5-C3 chemistry resulted in a total yield of 702 Mb (19× predicted average genome coverage), while the P4-C2 chemistry produced 1.7 Gb (46× average coverage) (see Table S1). However, the reads generated with P5-C3 chemistry had an average length of 8.3 kb, while the average read length with P4-C2 chemistry was 6.8 kb (see Table S1). Thus, while the P4-C2 chemistry generated more sequence output, the P5-C3 chemistry generated longer reads. Gap filling and scaffolding with PBJelly2 (version 14.9.9) (11), using long reads derived from both PacBio chemistries independently, significantly improved the genome assembly, resulting in a little over 300 contigs that exceeded 1 kb, while the longest contig reached >2.1 Mb (Table 1). The overall genome assembly improvement after gap filling is furthermore evidenced by the N_{50} value, which increased from 17.5 kb to approximately 650 kb (Table 1). However, although the contiguity of the genome assembly drastically increased upon the use of long reads for gap filling independent of the applied PacBio chemistry, the final assembly still contained a large amount of unknown-nucleotide sequences (Table 1).

In order to assess whether a single-step assembly that combines short- and long-read data, rather than gap filling with long reads of a previously generated assembly based on short reads, would increase assembly quality and contiguity, we assembled the ge-

TABLE 2 *Verticillium dahliae* strain JR2 assemblies based on different amounts of PacBio long reads

Metric	Value for data set:						
	SMRT.4	SMRT.6	SMRT.8	SMRT.10	SMRT.12	SMRT.14	SMRT.18
No. of SMRT cells used	4	6	8	10	12	14	18
Coverage (*n*-fold)	46.4×	72.1×	96.1×	120×	143.7×	167.1×	248×
Contig metrics with HGAP[a]							
No. of contigs ≥0 bp	246	45	49	41	41	34	35
No. of contigs ≥1,000 bp	246	45	49	41	41	34	35
Largest contig (bp)	957,075	8,522,516	9,231,296	5,501,910	5,496,487	5,496,279	9,231,737
Total length (bp)	36,019,813	36,390,158	36,496,508	36,298,366	36,439,073	36,407,468	36,472,797
N_{50} (bp)	271,892	3,085,282	4,168,662	2,910,158	3,361,205	3,361,230	3,399,208
No. of Ns/100 kb[b]	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Contig metrics with MHAP[a]							
No. of contigs ≥0 bp	95	132	77	55	50	47	48
No. of contigs ≥1,000 bp	95	132	77	55	50	47	48
Largest contig (bp)	3,355,274	1,305,931	3,215,544	3,814,805	5,484,470	4,267,138	5,486,069
Total length (bp)	36,785,530	35,897,226	36,545,821	36,523,003	36,589,360	36,382,335	36,635,502
N_{50} (bp)	1,816,396	567,445	1,167,265	2,569,351	3,068,688	2,330,944	3,358,862
No. of Ns/100 kb	0.00	0.00	0.00	0.00	0.00	0.00	0.00

[a] Software used for genome assemblies.
[b] Ns, unknown nucleotides.

nome of *V. dahliae* strain JR2 based on the previously generated PE and MP reads (see Table S1 in the supplemental material) (24) in combination with either of the two PacBio data sets (see Table S1) using SPAdes (version 3.0.0) (25). The assembly based on the P4-C2 data set contained 2,335 contigs with an N_{50} of ~47 kb and 659 scaffolds of >1 kb (Table 1). Interestingly, similar assembly statistics were obtained with the P5-C3 data set, showing that the differences in yield and read length obtained with the two chemistries did not affect the assembly process (Table 1).

Single-step assembly software packages were typically developed to assemble short reads only and are therefore not optimized to utilize long reads for gap filling. An alternative methodology is a two-step *de novo* assembly approach in which short reads are assembled, after which gap filling is performed with long reads (20, 26). Thus, we assembled the short reads using the A5 pipeline (version 2014.01.13) (27), followed by gap filling and scaffolding using PBJelly2 (version 14.9.9) (11) with a relatively low (between 10× and 50×) average genome coverage of PacBio reads (Table 1; see also Table S1 in the supplemental material). This two-step approach generated approximately 400 contigs of >1 kb with a contig N_{50} of about 2.3 Mb, thus clearly outperforming the single-step assemblies (Table 1). The superior quality of the two-step genome assembly is even more evident after scaffolding (Table 1). The gap-filled VerdaJR2v1.5 genome has contig metrics comparable to those of the two-step assembly that did not include optical map data (Table 1), demonstrating that inclusion of optical mapping in the assembly strategy does not improve contiguity. However, when comparing scaffold metrics, the addition of the optical map data resulted in superior genome assembly quality through the placement of more contigs into scaffolds. Nevertheless, the final assembly still contains a considerable amount of gaps (Table 1).

***De novo* genome assembly based on long reads only.** In order to assess whether we could assemble the *V. dahliae* JR2 genome based on long reads only, additional PacBio sequencing was performed using the P4-C2 chemistry. In total, 14 SMRT cells were used, resulting in 6 Gb of sequence (167× average genome coverage). This data set was randomly sampled and used to form subsets representing sequencing results from 4 (46×; SMRT.4), 6 (72×; SMRT.6), 8 (96×; SMRT.8), 10 (120×; SMRT.10), or 12 (144×; SMRT.12) SMRT cells, and all data sets were assembled using HGAP (version 2.0), as well as MHAP (version 1.5b1) software (see Table S2 in the supplemental material) (12, 28). The poorest assembly was obtained with HGAP (version 2.0) based on four SMRT cells, resulting in 246 contigs with an N_{50} of less than 0.3 Mb and a largest contig of less than 1 Mb (Table 2). However, all assemblies based on six or more SMRT cells generated comparable assembly outputs, with a total assembly size of ~36.5 Mb composed of up to 49 contigs, an N_{50} that exceeded 2.9 Mb, and a largest contig exceeding 5.5 Mb in all cases (Table 2). The fewest contigs, namely, 34, were produced based on 14 SMRT cells.

In order to determine its quality, the genome assembly based on 14 SMRT cells was aligned to the previously generated optical map (24), revealing that only a single contig was wrongly assembled (see Fig. S1 in the supplemental material). Nevertheless, all assemblies based on PacBio sequencing outperformed the hybrid assemblies as long as the sequencing depth exceeded 72× (Tables 1 and 2).

Assembly of a gapless genome of *V. dahliae* strain JR2. In an attempt to generate a gapless genome assembly of *V. dahliae* strain JR2, we generated a new assembly using HGAP software (version 2.0) (12) solely based on PacBio reads derived from two types of chemistries, comprising 2.7 million reads equivalent to 8.9 Gb (~248× coverage; 18 SMRT cells) (see Table S3 in the supplemental material). The assembly comprised 35 contigs, with the longest contig being ~9 Mb and an N_{50} of 3.4 Mb (Table 2). We subsequently aligned the contigs to the previously generated optical map (22), revealing five contigs that represented complete chromosomes (chromosomes 1, 3, and 6 to 8). The remaining three chromosomes could be assembled by ordering 12 contigs based on the optical map, followed by gap filling using PBJelly (version 14.9.9) (11). The gap-filled sequences were polished for errors that

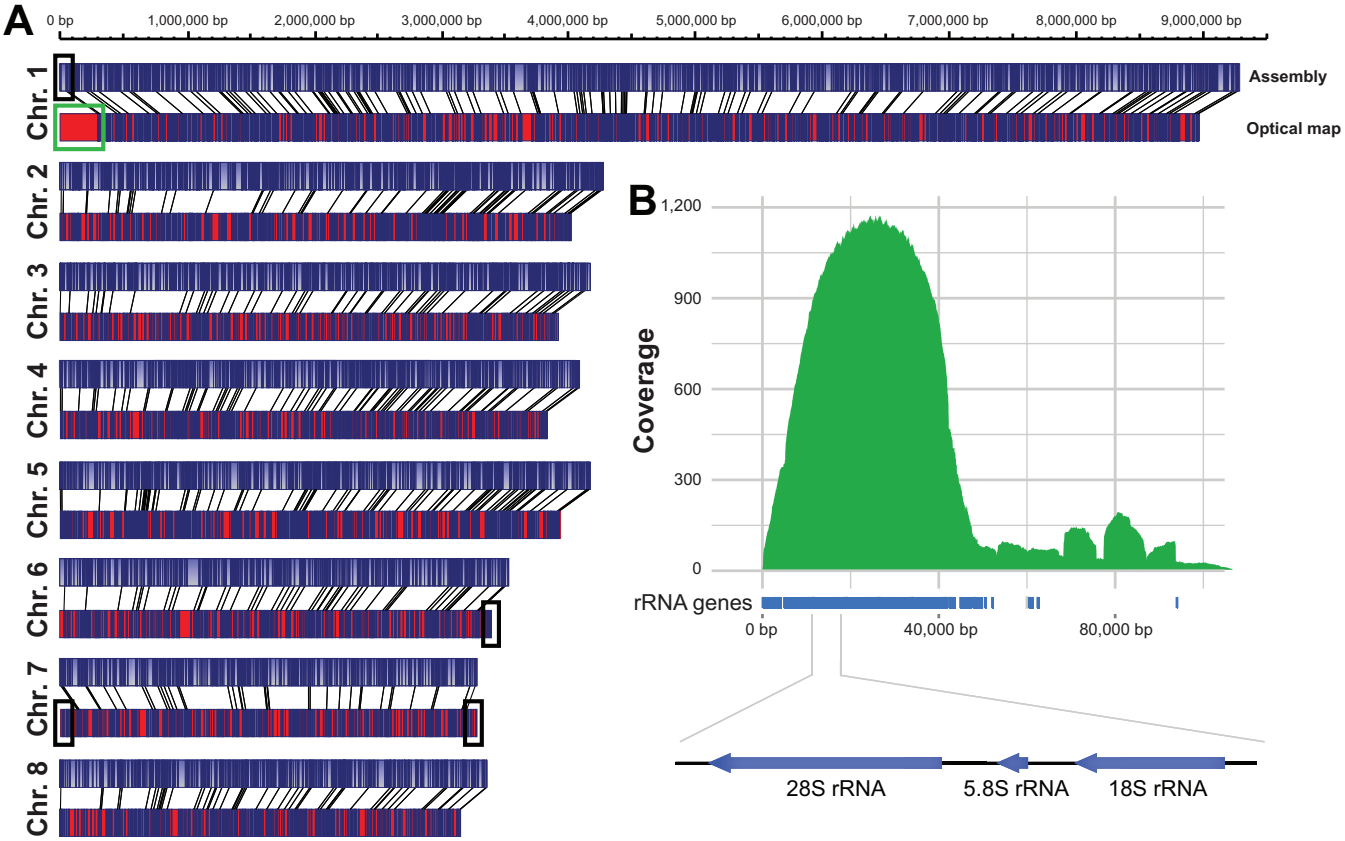

FIG 1 A gapless genome assembly of *Verticillium dahliae* strain JR2. (A) Alignment of the gapless genome assembly of *V. dahliae* strain JR2 with the optical map displays nearly perfect agreement. Represented in blue with blue lines is the genome assembly, while the optical map is represented in red with blue lines. Each blue line represents an NheI restriction site. Black lines represent alignments between the assembly and the optical map. Indicated in black and green boxes are length discrepancies between the assembly and the optical map due to the collapse of repetitive elements in the assembly. (B) Data for rRNA gene cluster located on the distal end of chromosome 1 (see green box in panel A). Local high read coverage (>1,000×) compared with the genomewide average of 15× coverage indicates the collapse of this region during the genome assembly. A single repeat unit of the *V. dahliae* rRNA gene is displayed, and its location in the assembly is marked.

could have been introduced by PBJelly2 (11) using Quiver (12). Thus, 17 of the 35 contigs obtained in the assembly spanned 98.1% of the predicted genome size in eight DNA molecules of contiguous sequence (Table 1), displaying perfect alignment to the optical map except for one edge of chromosome 1, one edge of chromosome 6, and both edges of chromosome 7 (Fig. 1A). However, mapping of the PacBio reads onto the assembly using BLASR (29) revealed a particularly high read coverage at the edges of these chromosomes, indicative of the collapse of repetitive elements. To investigate this hypothesis, *de novo* annotation of repetitive elements was performed (Table 3), indeed revealing the location of repetitive elements at these regions with high read coverage. More particularly, the high coverage at the edge of chromosome 1 could be attributed to of the collapse of the 300-kb ribosomal DNA repeat region to 50 kb by the assembly software (Fig. 1B). Thus, the length discrepancies between the assembly and the optical map are most likely the result of repetitive element collapse during the assembly process, and we conclude that we obtained a complete and gapless genome assembly of *V. dahliae* strain JR2.

To further assess the quality of the genome assembly, the correct assembly of the telomeres that comprise the chromosomal ends was assessed. To this end, we investigated the presence of the typical telomeric fungal repeat (TTAGGG) at each of the ends of the assembled chromosomes. Simple repeat analysis identified the typical fungal telomeric repeat in multiple copies (between 7 and 19) on both edges of each of all eight DNA molecules. In conclusion, we assembled the complete genome of *V. dahliae* strain JR2 in eight chromosomes from telomere to telomere without gaps.

Quality assessment of genome assemblies. The quality of genome assemblies is correlated with the quality of the sequencing reads used to generate the assembly. While Sanger and Illumina sequencing produce high-quality reads (2), PacBio sequencing generates long reads of relatively low-quality (~1 to 2 errors per 100 bp) that can only be used for genome assemblies after error correction. To assess how sequencing errors affect genome assemblies generated by PacBio long reads, we used high-quality short reads (PE and MP Illumina data) derived from *V. dahliae* strain JR2 (30) and mapped them independently onto the genome assemblies, after which sequence variants were identified as a proxy for sequence errors. The smallest amount of errors was observed in the SMRT.18 assembly generated by HGAP (version 2.0) (12) (Table 4). Interestingly, with the exception of the SMRT.4 assembly, all assemblies generated by HGAP (version 2.0) (12) carried fewer sequencing errors than the hybrid assemblies and assemblies generated by MHAP (version 1.5b1) (28) (Table 4). Our data therefore indicate that sequence errors that are intrinsic to long-

TABLE 3 Summary of transposable elements and other types of repetitive elements identified in *V. dahliae* strains JR2 and VdLs17

Type of element[a]	Value for *V. dahliae* strain: JR2 No. in genome	JR2 Coverage (bp)[b]	JR2 Coverage (%)[c]	VdLs17 No. in genome	VdLs17 Coverage (bp)[b]	VdLs17 Coverage (%)[c]
TEs						
SINEs	15	665	0	16	811	0
LINEs	324	124,209	0.34	311	167,003	0.46
LTR elements	1,071	2,428,443	6.72	1,006	2,430,766	6.76
DNA elements	269	114,336	0.32	272	150,768	0.42
Unclassified	1,557	1,286,043	3.56	1,351	1,098,298	3.05
Summary of TEs		3,953,696	10.94		3,847,646	10.7
Other repeats						
Small RNA	125	22,942	0.06	114	18,050	0.05
Satellites	74	7,336	0.02	71	7,003	0.02
Simple repeats	10,210	423,998	1.17	10,208	424,918	1.18
Low complexity	832	40,802	0.11	835	40,340	0.11
Total amt of repeats		4,446,122	12.3		4,336,001	12.05

[a] TEs, transposable elements; SINEs, short interspersed elements; LINEs, long interspersed elements; LTR, long terminal repeat.
[b] Total bases matching the element.
[c] % of genome covered by the element.

read sequencing do not affect genome assemblies generated by HGAP (version 2.0) (12) as long as sufficient read depth is used (minimum SMRT.6, corresponding to 72× average genome coverage). In contrast, genome assemblies based on a low coverage of long reads, such as the hybrid assemblies and the assembly solely based on a low coverage of PacBio long reads (SMRT.4, corresponding to 46× average genome coverage), as well as genome assemblies that are generated using MHAP (version 1.5b1) (28),

TABLE 4 Statistics for the various genome assemblies of *Verticillium dahliae* strain JR2 generated using Quast software and using VDAG_JR2v4.0 as the reference genome

Assembly, software used[a]	Data source: Library PE	Library MP	PacBio chemistry P4-C2	PacBio chemistry P5-C3	Optical map	Avg PacBio coverage[b]	Scaffold metric (no. of instances): SNPs[c]	Misassemblies	Genes that are: Complete	Partial	Missing
VerdaJR2v1.5	X	X			X		1,146	493	10,855	570	5
VerdaJR2v1.5	X	X	X			46	988	544	11,271	158	1
VerdaJR2v1.5	X	X		X		19	1,018	511	11,266	160	4
SPAdes 3.0	X	X				46	636	251	10,982	447	1
SPAdes 3.0	X		X			19	775	223	10,959	463	8
A5	X	X				46	661	369	11,349	78	3
A5	X		X			19	768	365	11,344	81	5
HGAP											
SMRT.4			X			46	1,089	21	11,149	167	114
SMRT.6			X			72	283	16	11,429	1	0
SMRT.8			X			96	175	12	11,429	1	0
SMRT.10			X			120	160	18	11,424	1	5
SMRT.12			X			143	146	21	11,429	1	0
SMRT.14			X			167	75	5	11,430	0	0
SMRT.18			X	X		248	41	13	11,430	0	0
VDAG_JR2v4.0			X	X	X	248	113	0	11,430	0	0
MHAP											
SMRT.4			X			46	10,270	15	11,410	16	4
SMRT.6			X			72	15,683	12	11,256	81	93
SMRT.8			X			96	11,256	15	11,397	25	8
SMRT.10			X			120	8,521	12	11,411	13	6
SMRT.12			X			143	7,579	13	11,425	4	1
SMRT.14			X			167	6,535	14	11,405	12	13

[a] HGAP 2.0 (12) and MHAP 1.5b1 (28) were used to generate assemblies.
[b] Average genome coverage of the PacBio data set used for the assembly.
[c] SNPs, single-nucleotide polymorphisms.

which lacks a polishing step after the assembly, are characterized by much higher error rates (Table 4). In conclusion, an error-free assembly of a genome like that of *V. dahliae* can be obtained upon HGAP (version 2.0) (12) assembly of PacBio long reads with ~72× average genome coverage, guaranteeing sufficient coverage for error correction and high sequence contiguity.

To further assess the quality of the different genome assemblies, we performed Quast analyses (31) that use a reference genome assembly and annotation to identify potential misassemblies. Here, we used the final *V. dahliae* JR2 genome (VDAG_JR2v4.0) assembly and annotation as a gold standard to investigate the quality of the other genome assemblies, as this assembly has been verified using the optical map. All genome assemblies generated based on long reads are characterized by a high level of contiguity, resulting in a high number of predicted genes (Tables 1, 2, and 4). However, for assemblies generated by MHAP (version 1.5b1) (28), a less-complete gene annotation was inferred compared with assemblies generated by HGAP (version 2.0) (12) (Table 4). Notably, all genome assemblies displayed misassemblies compared with the final genome assembly (Table 4). Therefore, the completeness and, thus, quality of the genome assembly are directly correlated with the amount of predicted genes in the assembly. Incomplete genome assemblies may therefore lead to a considerable underestimation of predicted genes and, thus, directly affect further biological studies.

Assembly of a gapless genome of *V. dahliae* strain VdLs17 using long reads and optical mapping. Different strains of *V. dahliae* are characterized by extensive structural rearrangements and chromosomal size polymorphisms (22). We previously showed that, despite a high degree of sequence identity, the genome of *V. dahliae* strain VdLs17 is structurally rearranged compared with that of *V. dahliae* strain JR2 (22). Thus, in order to evaluate the robustness of the approach described here, we made an attempt to assembled the genome of *V. dahliae* strain VdLs17 (23) based on PacBio reads and optical mapping. To this end, we generated about 300,000 reads equivalent to 1.6 Gb (~44× coverage) using P5-C3 chemistry (4 SMRT cells) (see Table S3 in the supplemental material). The assembly based on HGAP (version 2.0) (12) resulted in 119 contigs, with the longest contig being ~2.5 Mb and an N_{50} of 711 kb (Table 5). We subsequently aligned the contigs to the previously generated optical map of *V. dahliae* strain VdLs17 (23), revealing that about 98% of the genome was covered by the assembly. Surprisingly, this alignment revealed contig edges with considerable overlap that were not merged by the assembly software. Therefore, these overlapping edges were manually merged, gap filled with PBJelly (version 14.9.9) (11), and polished using Quiver (11, 12). This manual assembly produced eight gapless DNA molecules that matched the optical map almost perfectly (Table 5; see also Fig. S2 in the supplemental material). Similar to the genome assembly of *V. dahliae* strain JR2, an assembly collapse was identified at the edge of chromosome 1 where the ribosomal DNA cluster resides (see Fig. S2). Furthermore, two additional collapsed regions were identified on chromosome 4 (see Fig. S2). Similar to the *V. dahliae* strain JR2 assembly, telomeric repeats were found at both edges of each of the eight VdLs17 chromosomes. Thus, we also inferred a gapless genome assembly of *V. dahliae* strain VdLs17 in eight telomere-to-telomere chromosomes.

To determine the improvement of the gapless assembly generated here over the previously generated genome assembly based on Sanger sequencing and optical mapping (23), we compared both genome assemblies (Table 5). Surprisingly, a whole-genome alignment displayed a high number of inversions between the two assemblies (see Fig. S3 in the supplemental material). To resolve this observation, the previously generated genome assembly (23) was aligned to the optical map that had been used to generate this assembly. Unanticipated, a large number of assembly mistakes were revealed (see Fig. S4). Closer inspection of the assembly mistakes showed that, although the location of the placement of the scaffolds on the chromosomes was correct, the orientation of many of the scaffolds on the chromosomes was not (see Fig. S4). Thus, using long reads and optical mapping, we were able to identify and correct assembly mistakes in the previously generated and published VdLs17 genome assembly.

TABLE 5 Statistics of *Verticillium dahliae* strain VdLs17 genome assemblies

Metric	Value for assembly using:		
	Sanger sequencing + optical mapping[a]	PacBio only	PacBio + optical mapping
Contig metrics			
No. of contigs:			
≥0 bp	1,562	119	8
≥1,000 bp	1,525	118	8
Largest contig (bp)	216,594	2,545,020	6,210,300
Total length (bp)	32,902,348	36,288,516	35,973,870
N_{50} (bp)	43,309	711,766	5,894,008
No. of Ns/100 kb[b]	0	0	0
Scaffold metrics			
No. of contigs:			
≥0 bp	9	119	8
≥1,000 bp	9	118	8
Largest contig (bp)	6,048,892	2,545,020	6,210,300
Total length (bp)	36,874,636	36,288,516	35,973,870
N_{50} (bp)	4,180,501	711,766	5,894,008
No. of Ns/100 kb	10,770.33	0	0

[a] Genome assembly described in reference 23.
[b] Ns, unknown nucleotides.

Characterization of transposable elements in *V. dahliae* benefits from gapless genome. The biggest challenge in any genome assembly is the correct assembly of repetitive elements. Usually, relatively long repetitive elements, such as transposable elements (TEs), are poorly assembled in genome assemblies generated based on short reads, leading to an underestimation of the amount of TEs in the genome. However, TEs are important drivers of genome evolution (16) and, therefore, are relevant to many biological processes. For example, in the genomes of many plant-pathogenic fungi, TEs are found to accumulate at particular genomic regions, leading to genome plasticity that allows novel virulence factors called effectors to evolve (16). Moreover, TEs have been implicated in a phenomenon called "repeat-driven expansion," where the expansion of plastic genomic regions carrying highly variable genes with roles in pathogen virulence is mediate by TEs (32). Similar to those in the genomes of other fungal pathogens, TEs in the genome of *V. dahliae* are concentrated at distinct genomic locations and have been associated with genes that are important for virulence (23). Previous analyses have estimated that about 4% of the genomes of *V. dahliae* strains JR2 and VdLs17

is composed of repetitive elements (22, 23, 33). However, these estimations were likely compromised by the high level of fragmentation of these previous genome assemblies (23, 33). Using the gapless genome assemblies of *V. dahliae* strains JR2 and VdLs17, we now reinvestigated the amounts and types of TEs that can be found in these genomes. *De novo* TE prediction was performed on the *V. dahliae* strain JR2 genome, and about 20 TE families could be classified (see Table S4 in the supplemental material). Out of all of the TE families annotated, we identified 14 retrotransposon families and a few DNA transposon families. The retrotransposon families comprised one LINE (long interspersed element) retrotransposon family and 13 long-terminal-repeat (LTR) retrotransposon families that were further classified based on the presence of predicted open reading frames (ORFs) in the DNA sequence (34, 35). Thus, we identified seven retrotransposon families (VdLTRE1 to VdLTRE4, VdLTRE6, VdLTRE7, and VdLTRE12) displaying ORFs within the DNA sequence and four retrotransposon families (VdLTRE8 to VdLTRE11) lacking ORFs (see Table S4 in the supplemental material). Retrotransposons displaying ORFs in the DNA sequence are classified as autonomous elements, while retrotransposons lacking ORFs are considered nonautonomous elements. Interestingly, when we compared the TE repertoire of *V. dahliae* strain JR2 with that of strain VdLs17 that was previously assembled using Sanger sequencing (33), three families of autonomous LTRs (VdLTRE6, VdLTRE7, and VdLTRE12) and all nonautonomous LTR families (VdLTRE8 to VdLTRE11) found in JR2 were lacking in VdLs17, while one family (VdLTRE5) was only identified in VdLs17 and lacking in JR2 (33). However, when using the gapless VdLs17 genome assembly, all the TE families that were missing from previous analyses of the VdLs17 genome (33) were recovered, showing that the two strains have highly identical TE catalogues, which also agrees with their high levels (>99%) of overall genome nucleotide identity (22).

Finally, we reassessed the genomewide abundance of repetitive elements in the finished genome assemblies of *V. dahliae* strains JR2 and VdLs17, using the repertoire of nonredundant sequences generated by the combination of TE families identified in both genomes. LTR retrotransposons, in particular VdLTRE9 (see Table S4 in the supplemental material), are the most abundant TEs in both genomes (Table 3). Strikingly, in total, the repetitive elements in the *V. dahliae* genomes amount to 12% (Table 3), which is 3 times higher than all previous estimates for these genomes (22, 23, 33).

DISCUSSION

Advances in NGS technologies have allowed biologists to explore the genome sequences of a multitude of organisms across the tree of life to gain insight into their biology (20). However, only in a few cases has the massive amount of sequencing data that are typically obtained with these technologies been adequately transformed in to high-quality genome assemblies. Rather, highly fragmented assemblies that are biased toward genic regions have typically been obtained, while repetitive elements have remained significantly underrepresented. Considering the importance of repetitive elements for eukaryotic genome functioning and evolution, their correct and complete assembly is imperative for full understanding of the biology of an organism (36, 37). This is particularly relevant for filamentous pathogens, as transposable elements have been found to play crucial roles in the evolution of effector catalogues through repeat-driven expansion (16, 22, 26, 38). Fungal repetitive elements typically range in length between a few hundred base pairs and several kilobases (33, 39). Several sequencing and assembly strategies have recently been developed to assemble repeat-rich genomic sequences based on short reads (40, 41). However, these approaches are laborious and still challenging, and thus, their routine application may not be obvious. A way to improve assembly contiguity is by exploiting long reads that can span entire genomic repeats (20). Long-read technologies, such as SMRT and nanopore sequencing, can generate reads that span entire repetitive elements (6–10, 12, 42). Here, we show that the fragmented genome assemblies of the vascular wilt fungus *V. dahliae* that were previously obtained based on Sanger (23) and Illumina (22) sequencing could be significantly improved with the aid of long-read sequencing.

Our assembly results show that an average of ~20× genome coverage of long reads is sufficient for gap filling of the original *V. dahliae* strain JR2 genome assembly (Table 1), which was generated using a combination of short reads and optical mapping (22). Interestingly, the sequence contiguity of the gap-filled VerdaJR2v1.5 assembly was similar to that of the assembly based solely on long reads when using ~40× genome coverage (Table 2). This suggests that short reads do not contribute to the assembly quality if the depth of long reads is sufficient. Indeed, recent data on bacterial genomes show that genome assemblies obtained solely using long reads always outperform hybrid assemblies that include short reads, provided that the long-read sequencing depth is sufficient (~40 to 50×) (12, 42). However, when we compared genome assemblies generated by HGAP (version 2.0) (12) and MHAP (version 1.5b1) (28), it appeared that assemblies generated by HGAP (version 2.0) (12) carried fewer sequence errors (Table 4). The smaller amount of sequence errors in the HGAP (version 2.0) (12) assemblies is likely due to an extra step performed during the assembly procedure to polish the sequences and, thus, reduce the sequencing errors, which is lacking in MHAP (version 1.5b1) (28). Interestingly, however, MHAP (version 1.5b1) (28) produces fewer contigs than HGAP (version 2.0) (12) when using a long-read sequencing depth of ~40× (Table 2). These results confirm previous data from the MHAP developers indicating that MHAP (version 1.5b1) (28) produces fewer contigs than other software when only a low sequencing depth of long reads is used (28). The MHAP assembly results provide an appealing perspective for the application of long reads to larger genomes, using less data to achieve acceptable assembly statistics. Interestingly, the genome assembly quality did not increase with sequencing depths >72× (Table 2).

Although single repetitive elements and small clusters of repeats may be assembled based on long reads, the assembly of larger repeat clusters like those found in centromeres and in rRNA clusters remains challenging. Previously, optical mapping has been used successfully by us and others to order contigs into chromosomes for various bacterial (43), fungal (22, 44), plant (45, 46), and animal genome assemblies (14, 47). Therefore, the employment of optical mapping to order contigs on chromosomes that were not fully assembled by HGAP appeared to be a critical step to obtain a complete and gapless genome assembly of the two *V. dahliae* strains. Moreover, using optical mapping, we were able to correct mistakes that were generated by the assembly software at a relatively low sequencing depth (Fig. S1 in the supplemental material), as well as a large number of assembly mistakes in the genome assembly of strain VdLs17 that is publically available

through the Broad Institute repository (23). Comparison of the new, finished genome assemblies with the previously generated ones of *V. dahliae* strains VdLs17 and JR2 confirmed the assumption that the fragmentation of the original assemblies was mainly due to unassembled repetitive elements. Consequently, the amount of repetitive elements increased from 4% in the previous assemblies (22, 23) to 12% in the completed assemblies (Table 3). This finding will not only facilitate future studies of the role of particular repetitive elements in genome function and evolution but also allow studies addressing the *V. dahliae* genome structure, including, for instance, characterization of centromeric regions.

Nowadays, a prokaryotic *Escherichia coli* genome can be completely assembled based on sequencing of just a single SMRT cell (42). However, the complete assembly of eukaryotes is much more challenging, and only 10 gapless eukaryotic genome sequences have been reported thus far. Seven of these concern yeasts that have smaller genome sizes (<20 Mb) than filamentous fungi and a small amount of repetitive elements (~4%), while the remaining three concern filamentous fungi, namely, the dimorphic basidiomycete *Cryptococcus neoformans*, which has a relatively small genome (19 Mb) (48), and the ascomycetes *Myceliophthora thermophila* and *Thielavia terrestris*, which, like *V. dahliae*, belong to the class of *Sordariomycetes* and have larger genomes (37 to 39 Mb) (49). Importantly, the three filamentous fungal genome assemblies were generated based on Sanger sequencing, and for the two largest genomes, specific target sequencing of repetitive elements was used (49). Such an approach is laborious and expensive and not suitable for high-throughput eukaryotic genome sequencing in a routine manner. Thus, despite their relatively small genome sizes, still no complete, gapless genome assemblies have been reported for fungal organisms based on NGS technologies. Even the assembly of the fungal wheat pathogen *Zymoseptoria tritici* (formerly known as *Mycosphaerella graminicola*) still suffers from the presence of a few minor gaps despite the fact that this genome is generally portrayed as finished (50, 51). Nevertheless, it can be anticipated that the number of finished fungal genomes will increase significantly in the near future due to the application of long-read sequencing, as well as further dissemination of the use of optical mapping or the integration of sequence-based high-resolution genetic or physical maps. More recent technological advances have been able to automate optical mapping, as well as the subsequent imaging and data analysis. The resulting so-called whole-genome mapping can be executed in a high-throughput fashion and is therefore also suitable for more complex genome assembly projects (14). This, together with the ever-increasing read lengths due to improved sequencing chemistries, will bring gapless whole-genome assemblies within reach for complex eukaryotic genomes as well.

Conclusions. The complete genome assembly of two *V. dahliae* strains highlights the power of long-read DNA sequencing technology and establishes a standard for *de novo* genome assembly of haploid fungal genomes. We show that ~50× PacBio coverage is sufficient to achieve a high-quality genome assembly. However, an error-free assembly of a genome like that of *V. dahliae* can be obtained upon HGAP (version 2.0) (12) assembly of ~72× average genome coverage of PacBio long reads. Based on the combination of *de novo* assembly of long reads and optical mapping, we were able to assemble a gapless genome in a cost-effective manner. Technological advances in optical mapping, as well as in sequencing chemistries, will bring gapless whole-genome assemblies within reach for complex eukaryotic genomes as well. Importantly, finished genome assemblies will disclose genomic information that is imperative to fully appreciate an organism's biological complexity.

MATERIALS AND METHODS

DNA library preparation. The PacBio libraries were constructed using approximately 10 μg of genomic *V. dahliae* DNA that was mechanically sheared to a size of ~22 kb, using g-TUBES (Covaris, Inc., Woburn, MA) according the manufacturer's instructions. PacBio SMRTbell libraries were prepared by ligation of hairpin adaptors at both ends of the DNA fragment (52) using the PacBio DNA template preparation kit 2.0 (Pacific Biosciences of California, Inc., Menlo Park, CA) for SMRT sequencing on the PacBio RS II machine (Pacific Biosciences of California, Inc.) according to the manufacturer's instructions. Libraries were purified using Agencourt AMPure beads (Beckman Coulter, Inc., Brea, CA) to remove short inserts of <1.5 kb. Libraries were size selected using the BluePippin preparation system (Sage Science, Beverly, MA) with a minimum cutoff of 7 kb. The sheared DNA and final library were characterized for size distribution using an Agilent Bioanalyzer 2100 (Agilent Technology, Inc., Santa Clara, CA) along with a DNA12000 chip (Agilent Technology, Inc.).

DNA sequence data generation. PacBio sequencing data were generated at KeyGene N.V. (Wageningen, the Netherlands) using the PacBio RS II instrument. The DNA/polymerase binding kit 2.0 (Pacific Biosciences of California, Inc.) was used to anneal sequencing primers and DNA polymerase to the DNA fragments in order to make a complex for small-scale libraries, according to the manufacturer's recommendations. DNA template, polymerase, and primer complexes were diluted to a final concentration of 5 nM. The complexes were loaded onto the SMRT cells (Pacific Biosciences of California, Inc.). The sequencing kit 2.0 (Pacific Biosciences of California, Inc.) was used for sequencing using a 180-min sequence capture protocol along with stage start to maximize read length. Two different polymerases and chemistries were used to generate sequence reads for the *V. dahliae* strain JR2 genome. The P4 polymerase and C2 chemistry (P4-C2; Pacific Biosciences of California, Inc.) was applied to 14 SMRT cells, while the P5 polymerase and C3 chemistry (P5-C3; Pacific Biosciences of California, Inc.) was applied to four additional SMRT cells. The genome of *V. dahliae* strain VdLs17 was sequenced using four SMRT cells sequenced with P5-C3 chemistry.

Sequence assembly. Assemblies using SPAdes (version 3.0.0) (25) were performed using paired-end (PE) and mate-pair (MP) reads generated upon Illumina sequencing of short (500 bp) and long (5 kb) insert size libraries supplemented with long reads generated by either P4-C2 or P5-C3 PacBio chemistries. Gap filling of the previously generated VerdaJR2v1.5 genome assembly (22) was performed with PBJelly (version 14.9.9) (11), using long reads produced by either P4-C2 or P5-C3 PacBio chemistries. Our previously described hybrid assembly strategy (20) was used in combination with long reads produced with both PacBio chemistries.

Long-read assemblies were generated with HGAP (version 2.0) (12) and MHAP (version 1.5b1) (28) using default settings. Assembled sequences were aligned to the optical map with MapSolver (version 3.2) software (OpGen, Gaithersburg, MD) and manually ordered to compose chromosomes. Gap filling of the manually merged contigs was performed with PBJelly (version 14.9.9) (11) with default settings, except for the assembly stage, where the –maxTrim and –maxWiggle were set at 10,000 bp. Quiver (12) analysis was used to correct errors after gap filling.

Genome quality assessment. Telomeric repeats were identified by Tandem Repeats Finder (version 4.07b) (53). Repetitive elements were predicted using RepeatMasker (version 4.0.5) (54). An exhaustive and complete *Verticillium dahliae* TE database was generated by combining repetitive element sequences identified by LTRFinder (version 1.0.5) (55) and RepeatScout (version 1.0.5) (56), TEs previously identified in *V. dahliae* strain VdLs17 genome (33), and the RepBase database (57). Errors in the genome assembly generated based on long reads were iden-

tified by using Bowtie2 (58) to align previously generated high-quality short reads of *V. dahliae* strain JR2 (24) onto the genome assemblies that made use of the long reads and subsequent calling of sequence variants using FreeBayes (version 0.9.20) (59). Only sequence variants called with an accuracy higher than 99% and coverage of >10-fold were used in the analysis.

Accession numbers. Assembly data for the complete genomes can be found at the NCBI database under accession numbers GCA_000400815.2 and GCA_000952015.1 for *V. dahliae* strain JR2 and VdLs17, respectively.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00936-15/-/DCSupplemental.

Figure S1, PDF file, 0.8 MB.
Figure S2, PDF file, 0.5 MB.
Figure S3, PDF file, 0.4 MB.
Figure S4, PDF file, 0.7 MB.
Table S1, PDF file, 0.1 MB.
Table S2, DOCX file, 0.02 MB.
Table S3, DOCX file, 0.02 MB.
Table S4, DOCX file, 0.02 MB.

Acknowledgments

E.D., A.J., and A.H.J.W. are full-time employees of KeyGene N.V., a company offering next-generation sequencing services, including PacBio sequencing.

L.F. conceived of the study, participated in its design and coordination, performed analyses, and helped write the manuscript, M.F.S. participated in study design and coordination, performed analyses, and helped write the manuscript, E.D. generated assemblies for the different long-read data sets, G.C.M.V.D.B. produced the biological material used in the study, A.J. participated in the design of the study, A.H.J.W. participated in the design and coordination of the study, and B.P.H.J.T. conceived of the study, participated in its design and coordination, and helped write the manuscript. All authors read and approved the final manuscript.

REFERENCES

1. **Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER.** 2013. The next-generation sequencing revolution and its impact on genomics. Cell **155:**27–38. http://dx.doi.org/10.1016/j.cell.2013.09.006.
2. **Metzker ML.** 2010. Sequencing technologies—the next generation. Nat Rev Genet **11:**31–46. http://dx.doi.org/10.1038/nrg2626.
3. **Schatz MC, Delcher AL, Salzberg SL.** 2010. Assembly of large genomes using second-generation sequencing. Genome Res **20:**1165–1173. http://dx.doi.org/10.1101/gr.101360.109.
4. **Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, Nelson KE, Parkhill J, Pitluck S, Qin X, Read TD, Schmutz J, Sozhamannan S, Sterk P, Strausberg RL, Sutton G, Thomson NR, Tiedje JM, Weinstock G, Wollam A, Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium, Detter JC.** 2009. Genome project standards in a new era of sequencing. Science **326:**236–237.
5. **Treangen TJ, Salzberg SL.** 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet **13:**36–46. http://dx.doi.org/10.1038/nrg3117.
6. **Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J.** 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol **33:**296–300. http://dx.doi.org/10.1038/nbt.3103.
7. **Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, Craig JM, Langford KW, Samson JM, Daza R, Doering K, Shendure J, Gundlach JH.** 2014. Decoding long nanopore sequencing reads of natural DNA. Nat Biotechnol **32:**829–833. http://dx.doi.org/10.1038/nbt.2950.
8. **Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, Wilson RK, Turner SW, Korlach J, Eichler EE.** 2014. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res **24:**688–696. http://dx.doi.org/10.1101/gr.168450.113.
9. **Powers JG, Weigman VJ, Shu J, Pufky JM, Cox D, Hurban P.** 2013. Efficient and accurate whole genome assembly and methylome profiling of *E. coli*. BMC Genomics **14:**675. http://dx.doi.org/10.1186/1471-2164-14-675.
10. **Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin C-S, Rapicavoli NA, Rank DR, Li J, Catcheside DE, Celniker SE, Phillippy AM, Bergman CM, Landolin JM.** 2014. Long-read, whole-genome shotgun sequence data for five model organisms. Sci Data **1:**140045. http://dx.doi.org/10.1038/sdata.2014.45.
11. **English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA.** 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One **7:**e47768. http://dx.doi.org/10.1371/journal.pone.0047768.
12. **Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J.** 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods **10:**563–569. http://dx.doi.org/10.1038/nmeth.2474.
13. **Levy-Sakin M, Ebenstein Y.** 2013. Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. Curr Opin Biotechnol **24:**690–698. http://dx.doi.org/10.1016/j.copbio.2013.01.009.
14. **Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, Tosser-Klopp G, Wang J, Yang S, Liang J, Chen W, Chen J, Zeng P, Hou Y, Bian C, Pan S, Li Y, Liu X, Wang W, Servin B, Sayre B, Zhu B, Sweeney D, Moore R, Nie W, Shen Y, Zhao R, Zhang G, Li J, Faraut T, Womack J, Zhang Y, Kijas J, Cockett N, Xu X, Zhao S, Wang J, Wang W.** 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). Nat Biotechnol **31:**135–141. http://dx.doi.org/10.1038/nbt.2478.
15. **Shearer LA, Anderson LK, de Jong H, Smit S, Goicoechea JL, Roe BA, Hua A, Giovannoni JJ, Stack SM.** 2014. Fluorescence in situ hybridization and optical mapping to correct scaffold arrangement in the tomato genome. G3 (Bethesda) **4:**1395–1405. http://dx.doi.org/10.1534/g3.114.011197.
16. **Raffaele S, Kamoun S.** 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol **10:**417–430. http://dx.doi.org/10.1038/nrmicro2790.
17. **Bickhart DM, Liu GE.** 2014. The challenges and importance of structural variation detection in livestock. Front Genet **5:**37. http://dx.doi.org/10.3389/fgene.2014.00037.
18. **Seidl MF, Thomma BPHJ.** 2014. Sex or no sex: evolutionary adaptation occurs regardless. Bioessays **36:**335–345. http://dx.doi.org/10.1002/bies.201300155.
19. **Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE.** 2015. Resolving the complexity of the human genome using single-molecule sequencing. Nature **517:**608–611. http://dx.doi.org/10.1038/nature13907.
20. **Faino L, Thomma BPHJ.** 2014. Get your high-quality low-cost genome sequence. Trends Plant Sci **19:**288–291. http://dx.doi.org/10.1016/j.tplants.2014.02.003.
21. **Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, Radune D, Bergman NH, Phillippy AM.** 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol **14:**R101. http://dx.doi.org/10.1186/gb-2013-14-9-r101.
22. **de Jonge R, Bolton MD, Kombrink A, van den Berg GC, Yadeta KA, Thomma BPHJ.** 2013. Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. Genome Res **23:**1271–1282. http://dx.doi.org/10.1101/gr.152660.112.
23. **Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BPHJ, Chen Z, Henrissat B, Lee YH, 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 LJ.** 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.
24. **De Jonge R, van Esse HP, Maruthachalam K, Bolton MD, Santhanam P, Saber MK, Zhang Z, Usami T, Lievens B, Subbarao KV, Thomma BP.** 2012. Tomato immune receptor Ve1 recognizes effector of multiple fungal

pathogens uncovered by genome and RNA sequencing. Proc Natl Acad Sci U S A **109:**5110–5115. http://dx.doi.org/10.1073/pnas.1119623109.

25. **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.
26. **Seidl MF, Faino L, Shi-Kunne X, van den Berg GC, Bolton MD, Thomma BPHJ.** 2015. The genome of the saprophytic fungus *Verticillium* tricorpus reveals a complex effector repertoire resembling that of its pathogenic relatives. Mol Plant Microbe Interact **28:**362–373. http://dx.doi.org/10.1094/MPMI-06-14-0173-R.
27. **Tritt A, Eisen JA, Facciotti MT, Darling AE.** 2012. An integrated pipeline for de novo assembly of microbial genomes. PLoS One **7:**e42304. http://dx.doi.org/10.1371/journal.pone.0042304.
28. **Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM.** 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol **33:**623–630. http://dx.doi.org/10.1038/nbt.3238.
29. **Chaisson MJ, Tesler G.** 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics **13:**238. http://dx.doi.org/10.1186/1471-2105-13-238.
30. **Bulatovic M, Heijstek MW, van Dijkhuizen EHP, Wulffraat NM, Pluijm SMF, de Jonge R.** 2012. Prediction of clinical non-response to methotrexate treatment in juvenile idiopathic arthritis. Ann Rheum Dis **71:**1484–1489.
31. **Gurevich A, Saveliev V, Vyahhi N, Tesler G.** 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics **29:**1072–1075. http://dx.doi.org/10.1093/bioinformatics/btt086.
32. **Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, Ver Loren van Themaat E, Brown JK, Butcher SA, Gurr SJ, Lebrun MH, Ridout CJ, Schulze-Lefert P, Talbot NJ, Ahmadinejad N, Ametz C, Barton GR, Benjdia M, Bidzinski P, Bindschedler LV, Both M, Brewer MT, Cadle-Davidson L, Cadle-Davidson MM, Collemare J, Cramer R, Frenkel O, Godfrey D, Harriman J, Hoede C, King BC, Klages S, Kleemann J, Knoll D, Koti PS, Kreplak J, Lopez-Ruiz FJ, Lu X, Maekawa T, Mahanil S, Micali C, Milgroom MG, Montana G, Noir S, O'Connell RJ, Oberhaensli S, Parlange F, Pedersen C, Quesneville H, Reinhardt R, Rott M, et al.** 2010. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science **330:**1543–1546. http://dx.doi.org/10.1126/science.1194573.
33. **Amyotte SG, Tan X, Pennerman K, del Mar Jimenez-Gasco M, Klosterman SJ, Ma L-J, Dobinson KF, Veronese P.** 2012. Transposable elements in phytopathogenic Verticillium spp.: insights into genome evolution and inter- and intra-specific diversification. BMC Genomics **13:**314. http://dx.doi.org/10.1186/1471-2164-13-314.
34. **Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH.** 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet **8:**973–982. http://dx.doi.org/10.1038/nrg2165.
35. **Havecker ER, Gao X, Voytas DF.** 2004. The diversity of LTR retrotransposons. Genome Biol **5:**225. http://dx.doi.org/10.1186/gb-2004-5-6-225.
36. **Chénais B, Caruso A, Hiard S, Casse N.** 2012. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene **509:**7–15. http://dx.doi.org/10.1016/j.gene.2012.07.042.
37. **Pritham EJ.** 2009. Transposable elements and factors influencing their success in eukaryotes. J Hered **100:**648–655. http://dx.doi.org/10.1093/jhered/esp065.
38. **Grandaubert J, Lowe RG, Soyer JL, Schoch CL, Van de Wouw AP, Fudal I, Robbertse B, Lapalu N, Links MG, Ollivier B, Linglin J, Barbe V, Mangenot S, Cruaud C, Borhan H, Howlett BJ, Balesdent MH, Rouxel T.** 2014. Transposable element-assisted evolution and adaptation to host plant within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal pathogens. BMC Genomics **15:**891. http://dx.doi.org/10.1186/1471-2164-15-891.
39. **Daboussi MJ, Capy P.** 2003. Transposable elements in filamentous fungi. Annu Rev Microbiol **57:**275–299. http://dx.doi.org/10.1146/annurev.micro.57.030502.091029.
40. **McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS.** 2014. Illumina, TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoS One **9:**e106689.
41. **Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, Ishizuka KJ, Gissi C, Griggio F, Ben-Shlomo R, Corey DM, Penland L, White RA, III, Weissman IL, Quake SR.** 2013. The genome sequence of the colonial chordate, *Botryllus schlosseri*. Elife **2:**e00569. http://dx.doi.org/10.7554/eLife.00569.
42. **Liao YC, Lin SH, Lin HH.** 2015. Completing bacterial genome assemblies: strategy and performance comparisons. Sci Rep **5:**8747. http://dx.doi.org/10.1038/srep08747.
43. **Nagarajan N, Read TD, Pop M.** 2008. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. Bioinformatics **24:**1229–1235. http://dx.doi.org/10.1093/bioinformatics/btn102.
44. **Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B, Houterman PM, Kang S, Shim WB, Woloshuk C, Xie X, Xu JR, Antoniw J, Baker SE, Bluhm BH, Breakspear A, Brown DW, Butchko RA, Chapman S, Coulson R, Coutinho PM, Danchin EG, Diener A, Gale LR, Gardiner DM, Goff S, Hammond-Kosack KE, Hilburn K, Hua-Van A, Jonkers W, Kazan K, Kodira CD, Koehrsen M, Kumar L, Lee YH, Li L, Manners JM, Miranda-Saavedra D, Mukherjee M, Park G, Park J, Park SY, Proctor RH, Regev A, Ruiz-Roldan MC, Sain D, Sakthikumar S, et al.** 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature **464:**367–373. http://dx.doi.org/10.1038/nature08850.
45. **Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, Pape L, Mehan MR, Churas C, Pasternak S, Forrest DK, Wise R, Ware D, Wing RA, Waterman MS, Livny M, Schwartz DC.** 2009. A single molecule scaffold for the maize genome. PLoS Genet **5:**e1000711. http://dx.doi.org/10.1371/journal.pgen.1000711.
46. **Nowak MD, Russo G, Schlapbach R, Huu CN, Lenhard M, Conti E.** 2015. The draft genome of *Primula veris* yields insights into the molecular basis of heterostyly. Genome Biol **16:**12. http://dx.doi.org/10.1186/s13059-014-0567-z.
47. **Chamala S, Chanderbali AS, Der JP, Lan T, Walts B, Albert VA, dePamphilis CW, Leebens-Mack J, Rounsley S, Schuster SC, Wing RA, Xiao N, Moore R, Soltis PS, Soltis DE, Barbazuk WB.** 2013. Assembly and validation of the genome of the nonmodel basal angiosperm Amborella. Science **342:**1516–1517. http://dx.doi.org/10.1126/science.1241130.
48. **Janbon G, Ormerod KL, Paulet D, Byrnes EJ III, Yadav V, Chatterjee G, Mullapudi N, Hon CC, Billmyre RB, Brunel F, Bahn YS, Chen W, Chen Y, Chow EW, Coppée JY, Floyd-Averette A, Gaillardin C, Gerik KJ, Goldberg J, Gonzalez-Hilarion S, Gujja S, Hamlin JL, Hsueh YP, Ianiri G, Jones S, Kodira CD, Kozubowski L, Lam W, Marra M, Mesner LD, Mieczkowski PA, Moyrand F, Nielsen K, Proux C, Rossignol T, Schein JE, Sun S, Wollschlaeger C, Wood IA, Zeng Q, Neuveglise C, Newlon CS, Perfect JR, Lodge, Idnurm A, Stajich JE, Kronstad JW, Sanyal K, Heitman J, Fraser JA, Cuomo CA, Dietrich FS.** 2014. Analysis of the genome and transcriptome of *Cryptococcus neoformans* var. *grubii* reveals complex RNA expression and microevolution leading to virulence attenuation. PLoS Genet **10:**e1004261. http://dx.doi.org/10.1371/journal.pgen.1004261.
49. **Berka RM, Grigoriev IV, Otillar R, Salamov A, Grimwood J, Reid I, Ishmael N, John T, Darmond C, Moisan MC, Henrissat B, Coutinho PM, Lombard V, Natvig DO, Lindquist E, Schmutz J, Lucas S, Harris P, Powlowski J, Bellemare A, Taylor D, Butler G, de Vries RP, Allijn IE, van den Brink J, Ushinsky S, Storms R, Powell AJ, Paulsen IT, Elbourne LD, Baker SE, Magnuson J, Laboissiere S, Clutterbuck AJ, Martinez D, Wogulis M, de Leon AL, Rey MW, Tsang A.** 2011. Comparative genomic analysis of the thermophilic biomass-degrading fungi *Myceliophthora thermophila* and *Thielavia terrestris*. Nat Biotechnol **29:**922–927. http://dx.doi.org/10.1038/nbt.1976.
50. **Goodwin SB, M'Barek SB, Dhillon B, Wittenberg AH, Crane CF, Hane JK, Foster AJ, Van der Lee TA, Grimwood J, Aerts A, Antoniw J, Bailey A, Bluhm B, Bowler J, Bristow J, van der Burgt A, Canto-Canché B, Churchill AC, Conde-Ferràez L, Cools HJ, Coutinho PM, Csukai M, Dehal P, De Wit P, Donzelli B, van de Geest HC, van Ham RC, Hammond-Kosack KE, Henrissat B, Kilian A, Kobayashi AK, Koopmann E, Kourmpetis Y, Kuzniar A, Lindquist E, Lombard V, Maliepaard C, Martins N, Mehrabi R, Nap JP, Ponomarenko A, Rudd JJ, Salamov A, Schmutz J, Schouten HJ, Shapiro H, Stergiopoulos I, Torriani SF, Tu H, de Vries RP, Waalwijk C, et al.** 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella*

graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet **7:**e1002070. http://dx.doi.org/10.1371/journal.pgen.1002070.

51. **Dhillon B, Gill N, Hamelin RC, Goodwin SB.** 2014. The landscape of transposable elements in the finished genome of the fungal wheat pathogen *Mycosphaerella graminicola*. BMC Genomics **15:**1132. http://dx.doi.org/10.1186/1471-2164-15-1132.
52. **Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW.** 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res **38:**e159. http://dx.doi.org/10.1093/nar/gkq543.
53. **Benson G.** 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27:**573–580. http://dx.doi.org/10.1093/nar/27.2.573.
54. **Smit AFA, Hubley R, Green P.** 2006. RepeatMasker Open-3.0. 1996–2010. http://www.repeatmasker.org.
55. **Xu Z, Wang H.** 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res **35:**W265–W268. http://dx.doi.org/10.1093/nar/gkm286.
56. **Price AL, Jones NC, Pevzner PA.** 2005. *De novo* identification of repeat families in large genomes. Bioinformatics **21**(Suppl 1):i351–i358. http://dx.doi.org/10.1093/bioinformatics/bti1018.
57. **Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J.** 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res **110:**462–467. http://dx.doi.org/10.1159/000084979.
58. **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. http://dx.doi.org/10.1186/gb-2009-10-3-r25.
59. **Garrison E, Marth G.** 2012. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 (q-bio.GN). http://arxiv.org/abs/1207.3907v1.