Whole genome sequencing of seven Greek grapevine cultivars: genome assembly statistics

Vitis vinifera L. – QUAST assembly report

Dataset description

The genomes of seven Greek grapevine (*Vitis vinifera* L.) cultivars (Agiorgitiko, Assyrtiko, Malagousia, Limnio, Moschofilero, Xinomavro and Savvatiano) were sequenced using a hybrid approach that combined short reads (Illumina NovaSeq, 2 × 150 bp paired-end) with long reads (Oxford Nanopore Technologies), and the genomes were assembled de novo. Assembly quality was assessed with QUAST against the grapevine reference genome.

Species *Vitis vinifera* L.
Cultivars Agiorgitiko, Assyrtiko, Malagousia, Limnio, Moschofilero, Xinomavro, Savvatiano
Sequencing Illumina NovaSeq, 2 × 150 bp paired-end (short reads); Oxford Nanopore Technologies (long reads)
Quality assessment QUAST; reference genome length 486.3 Mb, GC 34.55%

Table 1. Genome assembly statistics of seven Greek grapevine cultivars (QUAST)

Metric	Agiorgitiko	Assyrtiko	Malagousia	Limnio	Moschofilero	Xinomavro	Savvatiano
Assembly statistics							
Assembly level	scaffolds	scaffolds	scaffolds	scaffolds	scaffolds	chromosomes	scaffolds
No. of sequences	25,754	32,585	25,187	5,475	17,900	33	27,419
Total length (Mb)	471.0	486.2	453.0	457.3	284.4	329.4	557.0
Largest sequence (Mb)	24.7	25.8	23.9	24.7	16.6	22.7	27.3
GC content (%)	35.66	35.82	35.96	34.98	36.28	34.90	34.30
N's per 100 kb	1,101.8	940.8	1,199.7	225.3	2,818.7	225.2	287.9
No. of sequences by minimum length							
≥ 0 bp	25,776	32,628	25,204	5,562	17,907	33	27,458
≥ 1 kb	25,725	32,551	25,166	5,406	17,891	33	27,372
≥ 5 kb	5,294	5,335	4,925	4,695	946	33	7,583
≥ 10 kb	1,474	1,664	1,381	2,838	110	33	3,104
≥ 25 kb	116	147	93	619	32	33	522
≥ 50 kb	36	37	35	116	32	32	85
Total length by minimum sequence length (Mb)							
≥ 0 bp	471.0	486.2	453.0	457.4	284.4	329.4	557.0
≥ 1 kb	470.9	486.2	452.9	457.3	284.4	329.4	556.9
≥ 5 kb	425.0	431.8	407.8	454.9	252.8	329.4	511.1
≥ 10 kb	398.4	406.1	383.2	441.3	247.4	329.4	479.5
≥ 25 kb	379.4	384.6	365.2	406.7	246.4	329.4	441.0
≥ 50 kb	377.0	381.1	363.4	390.2	246.4	329.4	426.8
Assembly contiguity							
N50 (Mb)	17.31	18.04	17.03	17.88	11.90	15.64	19.52
NG50 (Mb)	16.26	18.04	16.02	17.86	1.49	14.10	20.10
N75 (Mb)	14.53	14.45	14.42	14.46	9.17	14.10	1.54
NG75 (Mb)	14.53	14.45	0.03	13.78	–	–	16.32

Metric	Agiorgitiko	Assyrtiko	Malagousia	Limnio	Moschofilero	Xinomavro	Savvatiano
L50	12	12	12	11	11	10	13
LG50	13	12	13	12	22	15	11
L75	20	20	20	18	18	15	23
LG75	20	20	74	20	–	–	18
Comparison with the reference genome							
Genome fraction (%)	73.7	74.3	73.0	63.7	50.6	52.3	81.1
Duplication ratio	1.18	1.18	1.15	1.33	1.09	1.20	1.28
Total aligned length (Mb)	401.5	405.5	386.7	394.0	249.8	291.1	480.9
Largest alignment (kb)	640.6	963.3	936.5	394.0	552.8	324.5	715.6
Unaligned length (Mb)	61.4	73.3	58.2	59.3	25.8	35.1	69.7
Unaligned sequences (full + partial)	1463 + 8608 p	9349 + 8359 p	1472 + 8512 p	94 + 3694 p	2902 + 4563 p	0 + 33 p	855 + 9387 p
Mismatches per 100 kb	1,575.2	1,621.3	1,511.3	2,414.8	1,223.0	1,955.4	1,733.7
Indels per 100 kb	163.4	169.9	152.2	507.1	123.1	383.9	202.7
NA50 (kb)	31.2	29.0	33.3	16.8	39.7	19.3	31.1
NGA50 (kb)	28.7	29.0	27.4	14.4	1.1	6.2	40.9
NA75 (kb)	3.6	2.8	3.6	4.2	6.9	5.8	4.1
NGA75 (kb)	2.5	2.8	1.5	2.5	–	–	12.0
LA50	3,160	3,438	2,880	5,884	1,661	4,138	3,765
LGA50	3,416	3,439	3,431	6,814	18,197	11,137	2,775
LA75	14,426	17,453	13,438	19,958	5,614	11,791	15,886
LGA75	18,250	17,464	24,231	26,553	–	–	8,183
Structural differences from the reference (misassemblies)							
Misassemblies	30,253	31,832	28,819	46,144	15,685	33,766	36,611
Misassembled sequences	6,644	6,454	6,443	3,356	3,528	33	8,424
Length of misassembled sequences (Mb)	404.4	410.5	389.5	436.6	255.4	329.4	473.2
Local misassemblies	31,083	34,888	28,824	65,439	12,292	34,300	47,661
Scaffold gap ext. misassemblies	14,630	11,635	16,519	1,941	29,735	1,088	4,550
Scaffold gap loc. misassemblies	22,922	20,317	24,081	497	32,573	216	3,560
Unaligned misassembled sequences	4,263	4,049	4,108	803	1,735	0	4,328
Genomic features							
Genomic features (full + partial)	319347 + 17396 p	320700 + 16550 p	318245 + 17876 p	278164 + 17874 p	265208 + 38084 p	208973 + 15010 p	324905 + 13754 p
Predicted rRNA genes (full + partial)	122 + 7 p	109 + 9 p	104 + 9 p	105 + 6 p	18 + 3 p	157 + 7 p	71 + 5 p

N50/L50: length and number of the sequences that contain 50% of the assembly; NG50/LG50: the same relative to the reference genome length; NA50/NGA50: the same for aligned blocks. Duplication ratio > 1 indicates duplicated regions (e.g. uncollapsed haplotypes). – : not calculated (the assembly covers less than 75% of the reference). Misassemblies are counted against a reference genome from a different cultivar and largely reflect true structural differences between cultivars. p = partial.