Data in Brief 42 (2022) 108216

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Genomic data of two Greek *Vitis* varieties

George Tsiolas[a,*], Sofia Michailidou[a], Antiopi Tsoureki[a], Anagnostis Argiriou[a,b]

[a] *Institute of Applied Biosciences / CERTH, P.O. Box 60361, Thermi 57001, Thessaloniki, Greece*
[b] *Department of Food Science and Nutrition / University of the Aegean, Lemnos 81400, Greece*

ARTICLE INFO

Article history:
Received 21 February 2022
Revised 15 April 2022
Accepted 22 April 2022
Available online 27 April 2022

Dataset link: Greek Vitis (Original data)

Keywords:
Vitis cultivars
Whole-genome sequencing
Variant analysis
SNPs

ABSTRACT

The genetic material of *Vitis* varieties is crucial for the wine sector. In addition, genomic technologies applied in vitis germplasm characterization are important for the conservation of indigenous genetic reservoirs. Until recently the most common method to genetically identify vitis varieties was the use of Simple Sequence Repeats (SSR) along with SNP chips. Yet, with the progress in Next Generation Sequencing (NGS) technologies and the reduced sequencing cost per base, a twist in plant species genetic identification methods has occurred. Among them, the low coverage Whole-Genome Sequencing (lcWGS) method with downstream bioinformatic analysis for variant discovery and phylogenetic characterization is gaining scientific attention. In this dataset, shotgun sequencing data of two different Greek *Vitis* varieties, 'Razaki' and 'Vlachiko' are presented. Vitis cultivars were collected from the Aristotle University of Thessaloniki's (AUTH) ampelographic collection and have been previously phenotypically and genetically characterized. WGS libraries were sequenced on an Illumina® NovaSeq 6000 platform with the Illumina® NovaSeq 6000 S2 Reagent Kit (300 cycles). Raw sequence data used for analysis are available in NCBI under the Sequence Read Archive (SRA), with BioProject ID PRJNA805368. Reads were aligned to the reference genome of *Vitis vinifera* available from the EnsemblPlants database and formal analysis was conducted with the Genome Analysis Toolkit 4 (GATK4) pipeline. Data can be used to enrich our knowledge related to the genetic background of vitis cultivars

[*] Corresponding author.
E-mail address: george.tsiolas@certh.gr (G. Tsiolas).

https://doi.org/10.1016/j.dib.2022.108216

and can also serve as a threshold in the scientific community towards the construction of a genomic database of vitis cultivars.

Specifications Table

Subject	Biological sciences: Omics: Genomics
Specific subject area	Low coverage whole genome sequencing of two Greek vitis cultivars for cultivar identification and variant discovery
Type of data	Tables and Figures
How the data were acquired	WGS libraries were constructed using Illumina's Nextera DNA Flex library preparation kit. Sequencing was performed on an Illumina® NovaSeq 6000 platform using the Illumina® NovaSeq 6000 S2 Reagent Kit (300 cycles). The variant discovery was conducted using the Genome Analysis Toolkit 4 pipeline.
Data format	Raw and Analyzed
Description of data collection	Leaves from two grapevine varieties, 'Razaki' (white grape variety) and 'Vlachiko' (red grape variety), were obtained from the Ampelographic Collection of the Aristotle University of Thessaloniki.
Data source location	Institution: Institute of Applied Biosciences – Centre for Research and Technology Hellas City: Thessaloniki Country: Greece Latitude and longitude for analyzed data: 40.56806, 22.99713
Data accessibility	Repository name: NCBI SRA Data identification number: PRJNA805368 Direct URL to data: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA805368, https://www.ncbi.nlm.nih.gov/sra/?term=SRR17982062, https://www.ncbi.nlm.nih.gov/sra/?term=SRR17982063

Value of the Data

- Data add new knowledge on Vitis genetic variation at the level of variety.
- Data provides information on the genomic background of two Greek vitis varieties that can be used for future identification of unknown grapevine varieties.
- Viticulturists will benefit from results related to the functional characteristics of each variety through genomic selection.
- The data produced contribute to the preservation and the adoption of these vitis varieties in plant breeding schemes.

1. Data Description

Genomic sequencing data were generated with Illumina® NovaSeq 6000® platform using two paired-end libraries with insert size of approx. 300 bp. In total, 25.62 Gbases were generated with >Q30 of 98%; 12.91 Gbases for 'Razaki' variety and 12.71 Gbases for 'Vlachiko' variety (Table 1). Total coverage for each variety's genome was greater than 25x, which is sufficient to identify single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) across their genomes [1]. Raw sequencing data are available under the BioProject accession PRJNA805368 at NCBI's sequence read archive.

The majority of the discovered variants were SNPs with approximately 5.8×10^6 events in each variety and insertions/deletions with over 900×10^3 events per variety (Table 2). These

Table 1
Generated genomic data of Vitis varieties 'Razaki' and 'Vlachiko'.

BioSample	SRA Accession	Variety	Raw Reads	Gbases	Depth of coverage
SAMN25855225	SRR17982063	*Razaki*	47,292,258	12,917,903,048	26.57
SAMN25855226	SRR17982062	*Vlachiko*	47,404,592	12,714,283,858	26.15

Table 2
Type and number of variants per variety.

	Razaki		Vlachiko	
Summary Variant Statistics	InDels	SNPs	InDels	SNPs
Total number of loci	889,874	5,687,476	927,939	5,931,063
Number of variants (before filtering)	926,009	5,735,996	968,309	5,984,608
Number of variants processed (after filtering)	915,881	5,704,855	957,137	5,950,310
Number of multi-allelic variants (more than two alleles)	36,135	48,520	40,370	53,545
Number of effects	1,700,732	9,494,769	1,769,227	9,906,520
Reference genome total length	486,265,422	486,265,422	486,265,422	486,265,422
Reference genome effective length	486,265,422	486,265,422	486,265,422	486,265,422
Variant rate	1 every 530 bases	1 every 85 bases	1 every 508 bases	1 every 81 bases

numbers include SNPs and InDels found in unique sequences of the reference genome as well as in the repetitive genome fractions. The SNPs were primarily found in intergenic regions in contrast to the indels that were mainly found in intragenic regions causing frameshifts. Frameshift variants due to the indels are 6,847 in 'Razaki' and 7,024 in 'Vlachiko'. Missense variants due to the SNPs are 119,532 in 'Razaki' and 126,314 in 'Vlachiko'. SNPs and indels responsible for the gain of stop codons are 2,882 in 'Razaki' and 2,910 in 'Vlachiko'. The number and the type of variants of the affected Sequence Ontologies (SO) are presented in detail in Table 3 and Fig. 1.

2. Experimental Design, Materials and Methods

2.1. Sampling and library construction

Leaf tissues were obtained from two grapevine varieties 'Razaki' and 'Vlachiko', which are a part of the Ampelographic Collection of the Aristotle University of Thessaloniki. Leaves were ground to a fine powder in the presence of liquid nitrogen and subsequently, DNA extraction was conducted using the NucleoSpin Plant II kit (MACHEREY-NAGEL, Düren, Germany), according to the manufacturer's instructions. The quality of extracted DNA was assessed on a 0.8% agarose gel stained with 0.5 µg/ml ethidium bromide. DNA concentration was estimated by a fluorometric method on a Qubit 4.0 Fluorimeter using the Qubit® dsDNA BR assay kit (Invitrogen, Carlsbad, CA, USA).

Libraries were prepared with the Nextera DNA Flex library preparation kit following the manufacturer's instructions for an average insert size of 300 bp. Initially, libraries were quantified with the Qubit dsDNA BR kit and their average size was estimated by capillary fragment electrophoresis on a 5400 Fragment Analyzer system (Agilent Technologies, Santa Clara, CA, USA) using the DNF-477-0500 kit. Finally, library quantification was performed by qPCR using the KAPA Library Quantification kit for Illumina® sequencing platforms (Kapa Biosystems; Roche Diagnostics Corporation, Indianapolis, IN, USA) on a Rotor-Gene Q thermocycler (Qiagen, Hilden, Germany), and normalized in relation to their size. Libraries were sequenced on an Illumina® NovaSeq 6000® platform using the NovaSeq 6000 S2 Reagent Kit (300 cycles).

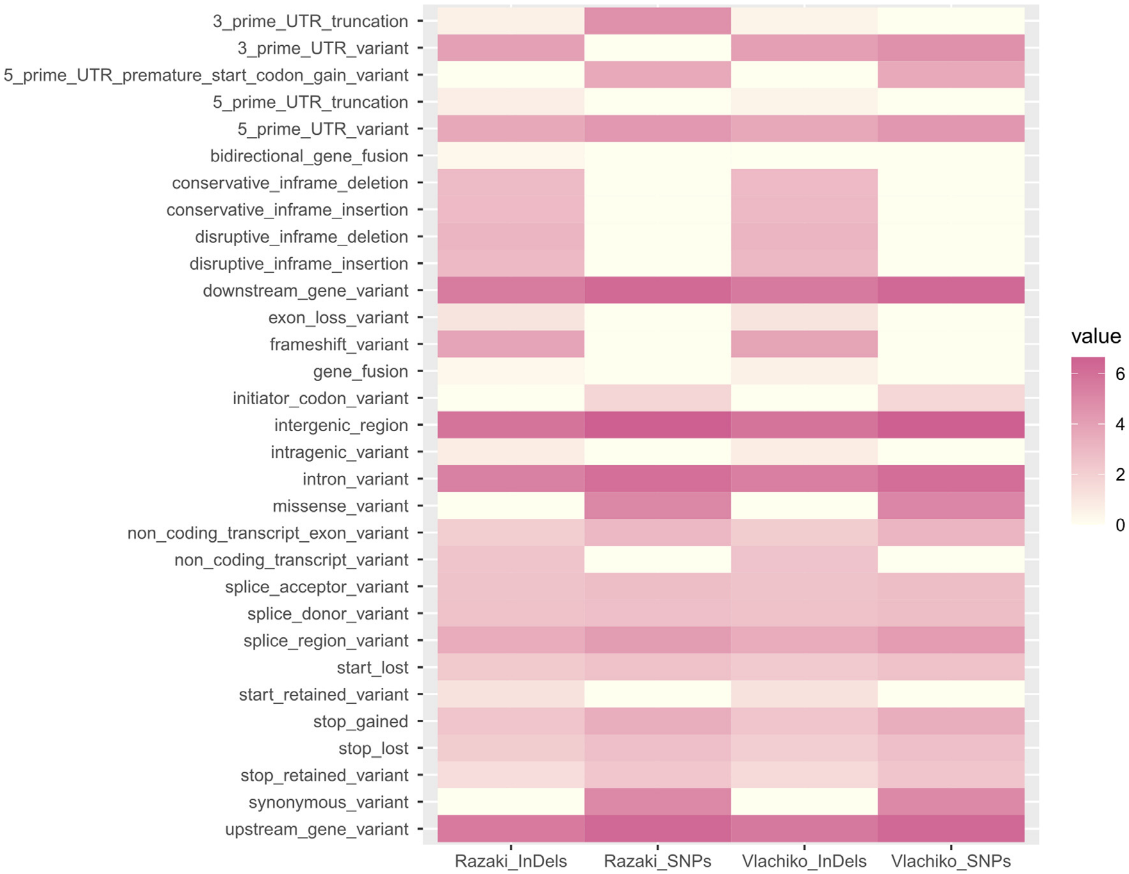

Fig. 1. Heatmap of variants for ‘Razaki’ and ‘Vlachiko’ varieties. Rows depict the affected Sequence Ontologies and columns the SNPs and InDels for each variety. Color scale refers to log10[(variants)+1].

Table 3
The number of variants and the affected Sequence Ontologies (SO).

Sequence Ontologies (SO) affected	Razaki		Vlachiko	
	InDels	SNPs	InDels	SNPs
3_prime_UTR_truncation	3	44,682	2	0
3_prime_UTR_variant	10,404	0	11,375	48,289
5_prime_UTR_premature_start_codon_gain_variant	0	3,776	0	3,889
5_prime_UTR_truncation	4	0	2	0
5_prime_UTR_variant	4,571	23,523	4,792	24,673
bidirectional_gene_fusion	1	0	0	0
conservative_inframe_deletion	786	0	794	0
conservative_inframe_insertion	807	0	907	0
disruptive_inframe_deletion	1,335	0	1,391	0
disruptive_inframe_insertion	952	0	1,012	0
downstream_gene_variant	371,034	1,811,783	386,922	1,906,153
exon_loss_variant	13	0	13	0
frameshift_variant	6,847	0	7,024	0
gene_fusion	1	0	3	0
initiator_codon_variant	0	51	0	50
intergenic_region	686,851	4,261,933	709,662	4,401,118
intragenic_variant	5	0	0	0
intron_variant	203,918	1,154,763	1,241,972	1,241,972
missense_variant	119,532	0	126,314	0
non_coding_transcript_exon_variant	968	122	1,216	133
non_coding_transcript_variant	0	307	0	331
splice_acceptor_variant	545	356	554	340
splice_donor_variant	522	428	570	422
splice_region_variant	14,691	3,220	16,125	3,346
start_lost	405	181	412	170
start_retained_variant	0	15	0	16
stop_gained	2,591	291	2,628	282
stop_lost	517	133	524	129
stop_retained_variant	231	25	253	35
synonymous_variant	95,902	0	103,131	0
upstream_gene_variant	1,973,306	412,606	2,045,032	423,890

2.2. Bioinformatics and data analysis

The quality of the reads was evaluated with the FastQC [2]. Raw reads were aligned to the reference genome of *Vitis vinifera* (12x) from EnsemblPlants (http://ftp.ensemblgenomes.org/pub/plants/release-52/fasta/vitis_vinifera/dna/) with MiniMap2 [3] and the command line options -x sr -a -R ‘@R\tID:<variety>\tLB:<variety>\tPL:ILLUMINA\tPM:NOVASEQ\tSM:<variety>’ without removing duplicate reads in this step. For variant discovery, Genome Analysis Toolkit 4 (GATK4) [4] pipeline was used. In detail, the duplicates duplicate reads were marked with the *MarkDuplicatesSpark* and the variants were recalibrated with the *BaseRecalibrator* using the filtered variants. The first round of variant discovery performed with *HaplotypeCaller*. Identified variants were filtered with *VariantFiltration* in order to filter out the variants with values of QD<2.0, FS>60.0, MQ<40.0, SOR>4.0, MQRankSum<-12.5 and ReadPosRankSum<-8.0. Final variants were obtained after the filtration of technical variants with the BaseRecalibrator and ApplyBSQR tools. The annotation of the final variants was performed with SnpEff [5].

Ethics Statements

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

CRediT Author Statement

George Tsiolas: Methodology, Analysis, Writing and Editing **Sofia Michailidou:** Methodology, Writing, Review and Editing. **Antiopi Tsoureki:** Analysis. **Anagnostis Argiriou:** Conceptualization, Review, Funding Acquisition, Project Administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Data Availability

Greek Vitis (Original data) (SRA NCBI).

Acknowledgments

This research received funding from the Emblematic Action Vitis Roads from the General Secretariat Research and Technology under grant number 2018ΣΕ01300000. Leaf tissues were kindly provided by Prof. N. Nikolaou, Head of the Ampelographic Collection of the Aristotle University of Thessaloniki.

References

[1] K. Song, L. Li, G. Zhang, Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology, Sci. Rep. 6 (2016) 1–7, doi:10.1038/srep35736.

[2] S. Andrew, FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]., (2015). http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[3] H. Li, Minimap2: Pairwise alignment for nucleotide sequences, Bioinformatics 34 (2018) 3094–3100, doi:10.1093/bioinformatics/bty191.

[4] G.A. Van der Auwera, M.O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, M.A. DePristo, From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline, 2013. doi:10.1002/0471250953.bi1110s43.

[5] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3, Fly (Austin) 6 (2012) 80–92, doi:10.4161/fly.19695.