ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/333302875

A genomic data mining pipeline for 15 species of the genus Olea

Article · May 2019

DOI: 10.14806/ej.24.0.922

CITATIONS
0

READS
32

8 authors, including:

Constantinos Salis
Agricultural University of Athens
4 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Eleni Papakonstantinou
Agricultural University of Athens
16 PUBLICATIONS **1** CITATION
SEE PROFILE

Katerina Pierouli
Agricultural University of Athens
4 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Lia Basdeki
Agricultural University of Athens
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

they both belong to my phd recearch View project

Frailsafe - EC Horizons 2020 View project

All content following this page was uploaded by Marianna Hagidimitriou on 03 June 2019.

The user has requested enhancement of the downloaded file.

EMBnet.journal
Bioinformatics in Action

Salis et al. (2019) EMBnet.journal 24, e922
http://dx.doi.org/10.14806/ej.24.0.922

Research Papers

A genomic data mining pipeline for 15 species of the genus *Olea*

Constantinos Salis[1], Eleni Papakonstantinou[1], Katerina Pierouli[1], Athanasios Mitsis[1], Lia Basdeki[1], Vasileios Megalooikonomou[2], Dimitrios Vlachakis[1,3,4]✉, Marianna Hagidimitriou[1]

[1]Laboratory of Genetics, Department of Biotechnology, School of Food, Biotechnology and Development, Agricultural University of Athens, Athens, Greece

[2]Computer Engineering and Informatics Department, School of Engineering, University of Patras, Patras, Greece

[3]Lab of Molecular Endocrinology, Center of Clinical, Experimental Surgery and Translational Research, Biomedical Research Foundation of the Academy of Athens, Athens, Greece

[4]Department of Informatics, Faculty of Natural and Mathematical Sciences, King's College London, London, United Kingdom

Competing interests: CS none; EP none; KP none; AM none; LB none; VM none; DV none; MH none

Abstract

In the big data era, conventional bioinformatics seems to fail in managing the full extent of the available genomic information. The current study is focused on olive tree species and the collection and analysis of genetic and genomic data, which are fragmented in various depositories. Extra virgin olive oil is classified as a medical food, due to nutraceutical benefits and its protective properties against cancer, cardiovascular diseases, age-related diseases, neurodegenerative disorders, and many other diseases. Extensive studies have reported the benefits of olive oil on human health. However, available data at the nucleotide sequence level are highly unstructured. Towards this aim, we describe an *in-silico* approach that combines methods from data mining and machine learning pipelines to ontology classification and semantic annotation. Fusing and analysing all available olive tree data is a step of uttermost importance in classifying and characterising the various cultivars, towards a comprehensive approach under the context of food safety and public health.

Introduction

The "Big Data" era is here and now. The amount of digitised data produced in modern society is increasing at an exponential rate and is estimated to account for five Tb (terabytes) for every human by 2020 (Egan 2013). Large-scale data is being generated each second in a wide range of areas, such as social networks, business and finance, and biosciences, posing a great challenge for data collection, storage, processing and analysis. In life sciences, the revolution following next-generation sequencing (Bahassi and Stambrook, 2014; Hui, 2014; van Dijk *et al.*, 2014) the Human Genome Project (Collins *et al.*, 2003; Green *et al.*, 2015), the advances in protein structure determination (Giege, 2013; Hekmat, 2015; Gavira, 2016), the development of biomedical and health informatics and of imaging informatics (Andreu-Perez *et al.*, 2015; Binder and Blettner, 2015) have inevitably led to an unprecedented data explosion. Consequently, biological data generated by genomics, proteomics, transcriptomics and metabolomics are characterised by a higher order complexity.

The advances in bioinformatics over the last decades has dramatically empowered researchers in handling omics information. An extensive set of computational tools, algorithms and databases have been developed for data analysis (Berger *et al.*, 2013). Still, at the rate at which data is generated and the ever increasing needs for storage, processing and meaningful analysis the spotlights are on the realm of bioinformatics. Moore's law predicts that computing power and storage capacity doubles every 15 years, whereas genomic data have grown tenfold every year since 2002 (Moore, 1965; Kahn, 2011). Storage space availability and computational power cannot keep up and fulfil the needs for rapidly expanding data-driven research domains (Papageorgiou *et al.*, 2018). Genomic raw data are not always useful as they come out from NGS and Illumina pipelines. The extraction, analysis and collection of data or the way they are annotated in databases, is far from to be standardised. Furthermore, genomic datasets are packed with noise or erroneous information (Fan *et al.*, 2014).

High-performance computing, smarter and faster algorithms and parallelisation for storage and processing seem to be the answer for data handling. As an example, column-oriented database systems have outmatched raw-oriented representation for data storage, enabling higher compressibility (Abadi *et al.*, 2009). Moreover,

Article history
Received: 17 January 2019
Accepted: 27 January 2019
Published: 22 May 2019

© 2019 Salis *et al.*; the authors have retained copyright and granted the Journal right of first publication; the work has been simultaneously released under a Creative Commons Attribution Licence, which allows others to share the work, while acknowledging the original authorship and initial publication in this Journal. The full licence notice is available at http://journal.embnet.org.

compressive algorithms have been developed which enable direct processing of the compressed data (Loh *et al.*, 2012; Berger *et al.*, 2016). Cloud computing infrastructures either driven by major software companies such as MS Azure and AWS, or joined multi-national initiatives, such as the Elixir[1] programme in Europe, strive towards the larger goal of unified and standardised metagenomics. The present study is focused on organising and mapping all available and dispersed olive tree nucleotide sequences to characterised regions on the reference genome of the recently published wild olive tree variant (*Olea europaea var sylvestris*) (Unver *et al.*, 2017).

The olive tree is one of the most ancient plants on earth and is primarily cultivated in the Mediterranean region which produces 90% of the olive oil consumed worldwide and controls almost 80% of the market share in exports (Bartolini and Petruccelli, 2002; Vasto *et al.*, 2014). Olive oil is the principal source of healthy fatty acids of the Mediterranean cuisine and is perceived as "superfood" rich in beneficial compounds (Vasto *et al.*, 2014; Gerber and Hoffman, 2015; Martinez-Gonzalez *et al.*, 2015). Extra virgin olive oil, rich in phenolic components, such as polyphenols (Barbaro *et al.*, 2014; Rigacci and Stefani, 2016), has been extensively studied for its antimicrobial, antioxidant and anti-inflammatory effect (Cicerale *et al.*, 2012). Additional to its nutraceutical benefits, the consumption of olive oil has been associated with reduced risk of various diseases, establishing it as a medical food. Indeed, many studies have denoted the protective effects of extra virgin olive oil for cardiovascular disease (Estruch *et al.*, 2006; Estruch *et al.*, 2013), diabetes (Salas-Salvado *et al.*, 2011, Salas-Salvado *et al.*, 2014), age-related and neurodegenerative diseases (Khalatbary 2013; Rodriguez-Morato *et al.*, 2015). Olive oil phenols have also been observed in several cancer cell lines to inhibiting proliferation and promote apoptosis, thus impending tumour aggregation.

The olive tree has been the subject of intensive research, whereas little is known about the phylogenetic relationships with other species. However, the molecular bases which conceal the differences between cultivars remain poorly understood. A resourceful pipeline for the analysis of the olive tree genetic and genomic information is essential towards the extraction of reliable conclusions about the molecular mechanisms of action of the olive tree and its beneficial effects on human health. On top of that, humanity will have to deal with the impact of climate change in the following years. Species in the plants' kingdom are profoundly affected, especially the olive tree, and climate change is posing a significant risk in olive cultivars. The potentiality to cultivate in different climate conditions, and expand in non-traditional continents, is highly dependent on the genetic profile of the species. The present study is an important precursor for handling and analysing raw genomics and genetics data from plants. The aim is to fill in the gaps in such analysis through filtering, clustering and classification with the use of ontology terms to discover the relational nodes of the available information. A data mining pipeline was performed on available genomic data of several species of the *Olea* genus, and we have developed an approach that may help to annotate plant genomic sequences better.

Methods

Data Collection

The dataset of genomic sequences was built by collecting data from the Nucleotide database of the NCBI. Keywords used for the retrieval and extraction of data were: "*Olea europaea*", "*europaea*", "protein", "dna", "nucleotide", "genome", "clone", "cultivar", "wild species", "propagating material", "subspecies", "*Oleaceae*", "olive", "gene", "protein" and "*Olea*". The analysis of the collected sequences was performed on three basic layers interacting with each other: the size of sequences, ontologies and nucleotide sequence similarities.

Data Filtering – First level of analysis

The dataset of nucleotide sequences obtained was filtered using the MATLAB platform and programming language. To reduce the noise, partial and variant sequences were removed from the dataset using a set of regular expressions. The new dataset, containing only full sequences, was then split into three sub-datasets by sequences' length as follow:

Group A: sequence length ≤ 1,000 bases
Group B: 10,000 bases ≤ sequence length ≥ 1000 bases
Group C: sequence length ≥ 10,000 bases

The dataset was split by sequence length because the goal was to isolate and focus on areas which correspond to protein sequences. As a result, Group A and Group B were used in the second level of analysis.

Data Mining and Semantic – Second level of analysis

Different groups of datasets were characterised by ontologies using clustering and classification algorithms. In this direction, the Bioinformatics Toolbox[2] was mainly employed for the computation, development, acquisition and modelling as a high-performance language for computing and programming, in a user-friendly operating environment (Cai, Smith *et al.*, 2005). In this direction, on the basis of the second level of analysis, a new database was created containing individual sub-datasets including: a) chloroplast, b) mitochondria, c) microsatellite, d) cultivars, e) protein, f) helicase, g) ATPase, h) plastid, i) trn gene, j) enzyme, k) species (Figure 1). Besides, the protein dataset was further categorised into smaller individual datasets, as follows: a) ribosomal, b) phosphatase, c) E3, d) FAR, e) Fbox, f) kinase, g) zinc-finger, h) pentatricopeptide.

[1]https://elixir-europe.org/

[2]https://www.mathworks.com/products/bioinfo.html

Salis et al. (2019) EMBnet.journal 24, e922
http://dx.doi.org/10.14806/ej.24.0.922

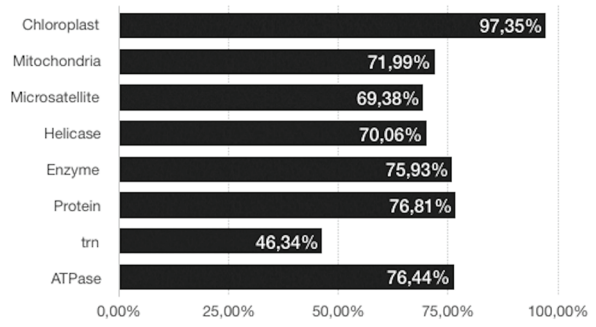

Figure 1. Percentage of *Olea europaea* nucleotide sequences in the individual sub-groups.

Analysis of genetic information – Third level of analysis

The third level of analysis consisted of grouping data obtained by the second analysis level by strict correlations of gene information. A classification function was created with the BLASTClust algorithm, in the Bio Linux operating system, to identify genetic similarity/dissimilarity between each genomic sequence BLASTClust inputs were nucleotide sequences that were analysed with the following parameters' values: coverage over 90% of the length of each sequence, with a 95% similarity cut off, and for the full-length (100% query cover) sequence, a 70% similarity cut off.

Results and Discussion

Data collection

During the data collection stage, we were able to put together more than 420000 nucleotide sequences from NCBI, which were then mapped on the genome of the wild olive tree, called "oleaster", which was assembled and annotated by Unver *et al.* (2017) (Unver, Wu *et al.*, 2017).

Data Filtering – First level of analysis

From the original pool of sequences, 8871% were classified as complete sequences and worthy of further investigation, while the remaining sequences (1129%) were partial or incomplete. The composition of the dataset was extremely heterogeneous; we indeed identified several genome regions of 15 different species of the genus *Olea* (Table 1), genome regions of 17 species of the plant kingdom and genome regions of 21 microorganisms, most of them affecting directly of indirectly the olive tree phenotype. From the filtered full-length sequences, 8674% referred to the genus *Olea* and particularly to the ontologies *europaea* and *oleaster*. In more detail, within the *Olea europaea* dataset were identified sequences with the ontologies "*europaea*", "*cuspidata*", "*laperrinei*", "*cerasiformis*", "*guanchika*" and "*maroccana*", which represent the subspecies of *Olea europaea* species, and 74% of the sequences were uncharacterised and represented as "*orphan*" sequences within *Olea europaea* species. In the remaining filtered data set, nucleotide sequences of several species of the genus *Olea* were identified, including *Olea exasperata*, *Olea capensis* with the ontologies "*hochstetteri*", "*macrocarpa*", "*enervis*", "*capensis*" and "*welwitschii*". In total, 72 cultivars were identified in the filtered dataset and another 20 cultivars discovered in the noise dataset with the partial and variant.

Regarding the split of the sequence pool by sequence length, 65,521% of the filtered sequences were in length Group A, 29,345% were in length Group B, while the rest 5,132% belonged to the length Group C.

Data mining and Semantics on Olea europaea - Second level of analysis

After the collection of all the available genetic and genomic information on the *Olea europea*, the possible relationships between the nucleotide sequences had to be identified. To this aim we needed to determine the integral nodes inside the selected dataset. Individual subgroups based on ontologies were filtered against the thousands of entries in Groups A and B. As an example,

Table 1. Ontologies per species identified in the dataset of the genus *Olea*

A/A	Species	Ontologies
1	*Olea europaea*	*Sylvestris*
		Europaea
		cuspidata/africana/indica/ferruginea
		Laperrinei
		cerasiformis
		guanchika
		maroccana
2	*Olea exasperata*	-
3	*Olea capensis*	*hochstetteri*
		macrocarpa
		enervis
		capensis
		welwitschii
4	*Olea lancea*	-
5	*Olea paniculata*	-
6	*Olea salicifolia*	-
7	*Olea rosea*	-
8	*Olea borneensis*	-
9	*Olea neriifolia*	-
10	*Olea brachiata*	-
11	*Olea javanica*	-
12	*Olea tsoongii*	-
13	*Olea schliebenii*	-
14	*Olea chimanimani*	-
15	*Olea woodiana*	*woodiana*

in the sub-dataset under the ontology "chloroplast", 1439 nucleotide sequences were clustered with a sequence average length of about ≈3841 bases 97,35% of the sequences referred to *Olea europaea*, while in the same group 12 species of the genus *Olea* were identified, three other species of the plant kingdom and four cultivars of the species *Olea europaea europaea*.

Similarly, the sub-dataset under the ontology "mitochondrial" contained 1439 sequences whose average length was about 1,241 bases. Among them, we identified two species of the genus *Olea*, 71,9% *Olea europaea* and 2,08% *Olea exasperata*, two other species of the plant kingdom and ten cultivars of the species *Olea europaea europaea*. Also, the sub-group under the ontology "micro satellite", contained 343 sequences with mean sequence length ~270 bases, was composed of 96,5% of *Olea europaea* sequences and the two subspecies, europaea and cuspidata. What is more, in the sub-group under the ontology "helicase", the 70,06% of sequences referred to *Olea europaea* and in the sub-dataset under the ontology "enzyme", *Olea europaea* sequences covered 74,27% of the dataset. The dataset under the ontology "trn" included 29 species, among which, two were *Olea europaea*, with six subspecies, and *Olea capensis*, with four subspecies.

On top of the above, in the dataset related to protein regions, 49 keywords with remarkable repeatability were identified, and 76,81% of the whole pool of sequences referred to *Olea europaea*, variant *Sylvestris*. Among the 49 keywords, eight became distinct and isolated from the dataset under the ontology "protein". The keyword "ribosomal", representing 15,27% of the dataset, "kinase" 30,06%, "E3" 9,85%, "phosphatase" 8,99%, "pentatricopeptide" 7,01%, "Fbox" 4,53% and "fatty acid- and retinol-binding protein (FAR)" 2,89% of the entire dataset, respectively. Lastly, based on the genetic information, we were able to identify nucleotide chains which bear the zinc-finger motif, representing the 8,67% of the protein dataset.

Analysis of genetic information – Third level of analysis

We were able to correlate nucleotide sequences of sub-datasets into clusters based on their genetic similarity. Clusters were made by considering that a cluster should have at least five nucleotide sequences with genetic similarity above the predefined threshold to be annotated as a cluster. Most of the sequences belonging to the length Group A, in the sub-datasets under the ontologies "trn" and "microsatellite", formed the highest number of clusters, seven and six respectively. The sub-group under the ontology "chloroplast" revealed four clusters and the sub-group under the ontology "mitochondrial" were all grouped in one cluster. The other sub-groups did not reveal any cluster. In total, the length Group A marked 19 clusters. Results showed that genomic sequences annotated as unknown sequences were clustered in an equal percentage as appeared in the initial dataset of *Olea* genus. Ultimately, the clustering results revealed that the hybrid pipeline could work well as a prediction tool.

Conclusions

This work represents the first attempt to cluster and identify olive tree cultivars based on their genome and genetic information. To date, cultivar assignment is done on the merit of morphology and pedigree in known breeds of the olive tree. However, since only recently the full genome of the *Olea europaea* variant was made public, it is now feasible to map on it all fragmented sequences of the olive tree genera and produce a set of genes or a gene panel that will be used to identify each cultivar with high accuracy genetically. Olive tree genetic fingerprinting holds great promise in the future for advanced control of olive tree breeding and olive oil that is consumed by the masses under the prism of food safety and public health.

Key Points

- Data mining and machine learning pipelines for the classification of olive tree cultivars.
- Olive tree genetic fingerprinting under the context of food safety and public health.
- Nutraceutical bioinformatics for olive oil as a medical food.

Acknowledgements

Research was supported by a Microsoft Azure for Genomics research Grant (CRM:0740983) and by the FrailSafe Project (H2020-PHC-21-2015 - 690140) "Sensing and predictive treatment of frailty and associated co-morbidities using advanced personalized models and advanced interventions", co-funded by the European Commission under the Horizon 2020 research and innovation program. EP was supported by the State Scholarships Foundation (IKY) - European Union (European Social Fund - ESF) and Greek national funds through the action entitled "Strengthening Human Resources Research Potential" via Doctorate Research in the framework of the Operational Program Human Resources Development Program, Education and Lifelong Learning of the National Strategic Reference Framework (NSRF) 2014 – 2020.

References

1. Abadi DJ, Boncz PA, Harizopoulos S (2009) Column-oriented database systems. Proceedings of the VLDB Endowment **2**(2): 1664-1665.
2. Andreu-Perez J, *et al.* (2015) Big data for health. IEEE J Biomed Health Inform **19**(4): 1193-1208, http://dx.doi.org/10.1109/JBHI.2015.245036
3. Bahassi el M and Stambrook PJ (2014) Next-generation sequencing technologies: breaking the sound barrier of human genetics. Mutagenesis **29**(5): 303-310. http://dx.doi.org/10.1093/mutage/geu031

4. Barbaro BG, *et al.* (2014) Effects of the olive-derived polyphenol oleuropein on human health. Int J Mol Sci **15**(10): 18508-18524. http://dx.doi.org/10.3390/ijms151018508
5. Bartolini G, Petruccelli R (2002) Classification, origin, diffusion and history of the olive. Food & Agriculture Org Book: ISBN-13: 978-9251048313
6. Berger BN, Daniels NM, Yu YW (2016) Computational Biology in the 21st Century: Scaling with Compressive Algorithms. Commun ACM **59**(8): 72-80. http://dx.doi.org/10.1145/2957324.
7. Berger B, Peng J, Singh M (2013) Computational solutions for omics data. Nat Rev Genet **14**(5): 333-346. http://dx.doi.org/10.1038/nrg3433.
8. Binder H, Blettner M (2015) Big data in medical science--a biostatistical view. Dtsch Arztebl Int **112**(9): 137-142. http://dx.doi.org/ 10.3238/arztebl.2015.0137.
9. Cai, J. J., Smith, D. K., Xia, X., & Yuen, K. Y. (2005). MBEToolbox: a MATLAB toolbox for sequence data analysis in molecular biology and evolution. BMC bioinformatics, **6**:64. http://dx.doi.org/10.1186/1471-2105-6-64
10. Cicerale S, Lucas LJ, Keast RS (2012) Antimicrobial, antioxidant and anti-inflammatory phenolic activities in extra virgin olive oil. Curr Opin Biotechnol **23**(2): 129-135. http://dx.doi.org/10.1016/j.copbio.2011.09.006.
11. Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: lessons from large-scale biology. Science **300**(5617): 286-290. http://dx.doi.org/10.1126/science.1084564
12. Cushman JC, Bohnert HJ (2000) Genomic approaches to plant stress tolerance. Curr Opin. Plant Biol. 3:117–124.
13. Egan BM (2013) Prediction of incident hypertension Health implications of data mining in the 'Big Data' era. J Hypertens **31**(11): 2123-2124. http://dx.doi.org/10.1097/HJH.0b013e328365b932.
14. Estruch R, Martinez-Gonzalez MA, Corella D *et al.* (2006) Effects of a Mediterranean-style diet on cardiovascular risk factors: a randomized trial. Ann Intern Med **145**(1): 1-11
15. Estruch R, *et al.* (2018) Primary prevention of cardiovascular disease with a Mediterranean diet. N N Engl J Med. **378**(25):e34. http://dx.doi.org/10.1056/NEJMoa1800389
16. Fan J, Han F, Liu H (2014) Challenges of Big Data Analysis. Natl Sci Rev **1**(2): 293-314. http://dx.doi.org/10.1093/nsr/nwt032
17. Gavira JA (2016) Current trends in protein crystallization. Arch Biochem Biophys **602**: 3-11. http://dx.doi.org/10.1016/j.abb.2015.12.010.
18. Moore GE (1965) Cramming more components onto integrated circuits. Electronics **38**(4).
19. Gerber M, Hoffman R (2015) The Mediterranean diet: health, science and society. Br J Nutr **113** Suppl 2: S4-10. http://dx.doi.org/101017/S0007114514003912.
20. Giege R (2013) A historical perspective on protein crystallization from 1840 to the present day. FEBS J **280**(24): 6456-6497. http://dx.doi.org/10.1111/febs.12580
21. Green ED, Watson JD, Collins FS (2015) Human Genome Project: Twenty-five years of big biology. Nature **526**(7571): 29-31. http://dx.doi.org/10.1038/526029a.
22. Hekmat D (2015) Large-scale crystallization of proteins for purification and formulation. Bioprocess Biosyst Eng **38**(7): 1209-1231. http://dx.doi.org/10.1007/s00449-015-1374-y.
23. Hui P (2014) Next generation sequencing: chemistry, technology and applications. Top Curr Chem **336**: 1-18. http://dx.doi.org/10.1007/128_2012_329.
24. Kahn SD (2011) On the future of genomic data. Science **331**(6018): 728-729. http://dx.doi.org/10.1126/science.1197891.
25. Khalatbary A R (2013) Olive oil phenols and neuroprotection. Nutr Neurosci **16**(6): 243-249. http://dx.doi.org/10.1179/1476830513Y.0000000052.
26. Loh PR, Baym, Berger B (2012) Compressive genomics. Nat Biotechnol **30**(7): 627-630. http://dx.doi.org/10.1038/nbt.2241.
27. Martinez-Gonzalez MA, *et al.* (2015) Benefits of the Mediterranean Diet: Insights From the PREDIMED Study. Prog Cardiovasc Dis **58**(1): 50-60. http://dx.doi.org/10.1016/j.pcad.2015.04.003.
28. Papageorgiou L, *et al.* (2018) Genomic big data hitting the storage bottleneck. EMBnet.journal **24**, e910. http://dx.doi.org/10.14806/ej.24.0.910
29. Ponti L, Gutierrez AP, Ruti PM, Dell'Aquila D (2014) Fine-scale ecological and economic assessment of climate change on olive in the Mediterranean Basin reveals winners and losers. Proc Natl Acad Sci U S A. **111**(15): 5598-5603. http://dx.doi.org/10.1073/pnas.1314437111.
30. Rigacci S, Stefani M (2016) Nutraceutical Properties of Olive Oil Polyphenols an Itinerary from Cultured Cells through Animal Models to Humans. Int J Mol Sci **17**(6). http://dx.doi.org/10.3390/ijms17060843.
31. Rodriguez-Morato J, Xicota L, Fito M, Farre M, Dierssen M, *et al.* (2015) Potential role of olive oil phenolic compounds in the prevention of neurodegenerative diseases. Molecules **20**(3): 4655-4680. http://dx.doi.org/10.3390/molecules20034655.
32. Salas-Salvado J, *et al.* (2011) Reduction in the incidence of type 2 diabetes with the Mediterranean diet: results of the PREDIMED-Reus nutrition intervention randomized trial. Diabetes Care **34**(1): 14-19. http://dx.doi.org/10.2337/dc10-1288.
33. Salas-Salvado J, Bullo N, Estruch R *et al.* (2014) Prevention of diabetes with Mediterranean diets: a subgroup analysis of a randomized trial. Ann Intern Med **160**(1): 1-10. http://dx.doi.org/10.7326/M13-1725.
34. Unver T, Wu Z, Sterck L *et al.* (2017) Genome of wild olive and the evolution of oil biosynthesis. Proc Natl Acad Sci U S A **114**(44): E9413-E9422. http://dx.doi.org/10.1073/pnas.1708621114.
35. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet **30**(9): 418-426. http://dx.doi.org/10.1016/j.tig.2014.07.001.
36. Vasto S, Barera A, Rizzo C, Di Carlo M, Caruso C, *et al.* (2014) Mediterranean diet and longevity: an example of nutraceuticals? Curr Vasc Pharmacol **12**(5): 735-738.

View publication stats