Departamento de Bioquímica y Biología Molecular

Genetic and genomic approaches to characterize crop varieties

Leticia Ayllón Egea

Tesis Doctoral

Córdoba

TITULO: *Genetic and genomic approaches to characterize crop varieties*

AUTOR: *Leticia Ayllón Egea*

Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

www.uco.es/publicaciones
publicaciones@uco.es

TESIS DOCTORAL

Genetic and genomic approaches to characterize crop varieties

Memoria de Tesis Doctoral presentada por Leticia Ayllón Egea, Licenciada en Biología, para optar al grado de Doctor por la Universidad de Córdoba con la mención de Doctorado Internacional.

Dirigido por:

Dr. Gabriel Dorado Pérez

Catedrático de Bioquímica y Biología Molecular

Dra. Pilar Hernández Molina

Científico Titular

Córdoba

TÍTULO DE LA TESIS: Genetic and genomic approaches to characterize crop varieties

DOCTORANDO/A: Leticia Ayllón Egea

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS
(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

La doctoranda Leticia Ayllón Egea

La doctoranda Leticia Ayllón Egea ha aprovechado satisfactoriamente su contrato predoctoral. Ha adquirido conocimientos teóricos y prácticos necesarios para la realización de la tesis doctoral. Asimismo, ha realizado las actividades de formación complementarias del programa de doctorado "Ingeniería Agraria, Alimentaria, Forestal y de Desarrollo Rural Sostenible" en el que se encuentra inscrita. La alumna ha impartido 15 horas de clases en el curso 2014/2015, en la asignatura Bioquímica Experimental I del Grado de Bioquímica, 12 horas en el curso 2015/2016, en la asignatura Bioquímica del Grado de Biología y 27 horas en el curso 2016/2017, en esta misma asignatura. Todas ellas dentro del Departamento de Bioquímica y Biología Molecular, de la Universidad de Córdoba. Por otro lado, ha realizado una estancia de investigación en el Departamento de Agricultura, Alimentación y Ambiente (Di3A) de la Universidad de Catania en Sicilia (Italia), para obtener la mención internacional en el doctorado. Ya se ha publicado un artículo científico derivado de la tesis doctoral (indicado abajo), y otros están en preparación:

Egea LA, Mérida-García R, Kilian A, Hernandez P, Dorado G (2017): Assessment of genetic diversity and structure of large garlic (*Allium sativum*) germplasm bank, by Diversity Arrays Technology "genotyping-by-sequencing" platform (DArTseq). Frontiers in Genetics 8: 98 (9 pp). DOI: 10.3389/fgene.2017.00098.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, a 5 de octubre de 2017

Firma del/de los director/es

Fdo.: Gabriel Dorado Pérez

Fdo.: Pilar Hernández Molina

ACKNOWLEDGEMENTS / AGRADECIMIENTOS / RINGRAZIAMENTI

Es así que otras aguas se presienten
azules, más allá, volviendo El Cabo,
y en los acantilados amanecen
palomas y zureos,
sirenas nuevas,
que desde el farallón de la esperanza
pueblan el aire.

Javier Egea, Troppo Mare.

Después de más o menos cuatro años, esta Tesis toca a su fin. Por fin llega el momento de agradecer a todas las personas que de un modo u otro han estado junto a mí en este tiempo. Probablemente me deje a muchos por el camino, pero espero que sean los menos posibles.

En primer lugar, quiero expresar mi gratitud a los doctores, Gabriel Dorado Pérez y Pilar Hernández Molina, por dirigir mi Tesis Doctoral.

Además, me gustaría dar las gracias a la Junta de Andalucía, concretamente a la Consejería de Economía y Conocimiento y a la Secretaría General de Universidades, Investigación y Tecnología, por la concesión del contrato predoctoral que me ha permitido realizar este trabajo.

Igualmente, agradezco a Francisco Mansilla, Isabel María González Padilla, Marcelino de los Mozos Pascual, Antonio Escolano García, Jesús Martín Sánchez y Jaime Martín el facilitarme las muestras de ajo y la información necesaria para los análisis realizados. Así mismo, a Angjelina Belaj por su ayuda en la determinación y aclaración de las sinonimias de las variedades de olivo.

Al SCAI, en especial a Mercedes Cousinou Rodríguez y Laura Redondo Sánchez por su paciencia y dedicación.

Ringrazio i professori Alessandra Gentile, Stefano La Malfa, Gaetano Distefano e Angela Roberta Lo Piero, per avermi accolta all'Università di Catania durante il mio soggiorno di ricerca. Grazie per tutte le attenzioni, gli insegnamenti e per avermi fatta sentire come a casa. Vorrei ringraziare anche il professore Alberto Continella e i ragazzi del Dipartimento di Agricoltura, Alimentazione e Ambiente, per tutti i bei momenti passati insieme. Francesco, amico, ho troppe cose di cui ringraziarti, non si possono descrivere con le parole. Grazie per esserci sempre e per la tua amicizia.

A mis compañeras de laboratorio, Rosa y Tere, por las muchas horas de trabajo juntas. A mis demás compañeros del IAS (fijo me dejo a alguien): los Álvaros, Carmen, Juanma, Kiki, Manolo, Nuria y Valle, entre otros, por las comidas, cervezas, escuchas y apoyos. A Barbara, Giuseppe y Ludovica, mis queridos italianos españoles, gracias por vuestra amistad y por vuestra ayuda con el italiano.

A toda la gente que he conocido en este ilusionante proyecto que hemos empezado, la Asociación de Investigadores de Córdoba, y que me han hecho sentirme menos perdida y más acogida en Rabanales. En especial a Adrián, Almudena, Aurora y Maribel.

No me puedo olvidar, por supuesto, de mi gente de bioquímica y allegados. Gracias a todos por vuestra mistad. A mis chicos de fresa, Javi y Félix, sin vosotros esto no habría salido para delante. A la demás gente de fresa, por su amabilidad. A Casi, por tener siempre tu apoyo, tanto en la tesis como en la asociación. A Andrés, compañero de tantas cosas. A mi familia adoptiva, en especial al doctor Jesús Valentín Jorrín Novo, nunca podré agradecerte lo que para mí a supuesto y significado tu acogida. Me has escuchado, apoyado, enseñado y ayudado muchísimo más de lo que podría haber esperado. A Cristina y Rosa, qué habría hecho sin vosotras dos, tampoco tengo forma de agradeceros todo lo que habéis hecho por mí. ¡Gracias, gracias, gracias! A la doctora Ana Maldonado Alconada, Conchi, Dani, Isa, Kamilla, al doctor Manuel Rodríguez Ortega, Mari Carmen, Patricia, Víctor y todos los que hayan pasado por el grupo de Jesús, aunque haya sido por poco tiempo, por hacerme sentir una más.

También quería agradecer a los doctores Juan Jurado Carpio y Juan Muñoz Blanco, por haberme permitido impartir docencia en las asignaturas coordinadas por ellos. Ha sido una experiencia muy enriquecedora y gratificante. Igualmente, quería reconocer al doctor Conrado Moreno Vivián la ayuda prestada como director de departamento. Por último, a las doctoras Carmina Michán Doña, Josefa Muñoz Alamillo, Lara Sáez Melero, Leovigilda Ortiz Medina y Rosario Blanco Portales, por estar siempre pendientes de mí.

A mis amigas de siempre, Ale, Laura, Rosa y Teresa, por seguir ahí pese a la distancia y mi falta de tiempo libre.

A mi familia, por su amor y apoyo incondicional. En especial a Ahinara y Tista, por la compañía cordobesa. A Pablo, por sus consejos. A mi madre, Ángela, a quien tampoco sé cómo agradecerle una mínima parte de toda su paciencia, apoyo, ayuda y compresión.

A Rafa, para quien tampoco hay palabras suficientes para darle las gracias. Sin ti, esto habría sido mucho más difícil.

A todos, mi más sincero agradecimiento.

TABLE OF CONTENTS

TABLE OF CONTENTS

ABSTRACT / RESUMEN

Introduction

Plant-germplasm banks are biodiversity reservoirs. They can harbor varieties in many ways, like seeds, arboretums, seeding seasonal crops and maintaining *in vitro* cultures. Early germplasm-banks records belong to Egyptian and Babylonian societies. Probably, Nikolaj Ivanovič Vavilov was the first person that stressed the necessity of creating germplasm banks for society's welfare. Traditionally, varieties were stored according to morphological characteristics. Yet, due to phenotypical plasticity of plants, this classification could lead to synonymy and homonymy. This situation has triggered the necessity of characterizing germplasm banks not only by morphologic criteria, but also by molecular markers. The latter have significantly improved in recent years. Thus, they have evolved from peptide- to DNA-based methods. Moreover, the latter have greatly improved thanks to Polymerase Chain-Reaction (PCR), including recent high-throughput approaches. This has been possible thanks to the emergence of technologies like Second-Generation Sequencing (SGS) and Third-Generation Sequencing (TGS) platforms. Currently-used approaches for genotyping include Simple-Sequence-Repeat (SSR), Single-Nucleotide-Polymorphisms (SNP) and Genotyping-by-Sequencing (GBS). In summary, molecular characterization allows better genetic identification, understanding of biological functions and finding of genetic relationships for evolutionary biology. They are also used for plant conservation biology, biosecurity studies, and germplasm-bank management. Besides, they are excellent tools for assisted plant breeding, as well as intellectual-property certification and traceability applications.

Research content

In this Doctoral Thesis, garlic and olive-tree germplasm banks have been analyzed by molecular markers. In short, the former has been traditionally used worldwide as a common food ingredient and natural healing remedy in pharmacology/medicine. This is ought to its interesting beneficial attributes, reducing high blood pressure, cholesterol and atherosclerosis. Garlic also has preventive effects against cancer and antimicrobial activity. On the other hand, olive-tree importance is undeniable. It is one of the most cultivated species worldwide, especially in the Mediterranean basin, and it is the second species for oil production after palm oil. The total area of cultivation is ten million hectares. Its countless culinary and medicinal properties have enhanced its expansion to non-traditional producer and consumer areas.

The main purpose of garlic-germplasm bank characterization was to study genetic diversity and structure of 417 samples from the main Garlic-Germplasm Bank at "Instituto Andaluz de Investigación y Formación Agraria, Pesquera, Alimentaria y de la Producción Ecológica" (IFAPA) of "Junta de Andalucía", Cordoba University, and "Centro de Ensayos de Evaluación de Variedades" at "Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria" (INIA) in Madrid. The chosen technique was DArTseq (Diversity Arrays Technology), which allows sample characterization in species without reference genome or other previous genetic information. This way, a core collection was created in order to reduce the number of original accessions in the bank, without losing genetic diversity. In addition, polymorphic sequences of garlic generated in DArTseq analyses were used to ascertain their identities, functions, Gene Ontology (GO) terms, and metabolic pathways by looking for identities in other plant-species databases.

On the other hand, the main objective of the olive-tree chapter was to describe a "closed-tube" and cost-effective method to genotype varieties when previous genetic information is available. In this case, 83 samples were analyzed using six molecular markers and High-Resolution Melting (HRM) technique. Additionally, although a low number of markers was used, characterization analyses were found to be in agreement to previous works.

Conclusion

Both molecular-marker techniques (DArTseq and HRM) showed consistent-genotyping results according to prior passport data. Garlic germplasm-bank size was significantly reduced, which indicates that DArTseq analysis is a suitable technology for high-throughput genotyping without available genetic information. To our knowledge, this is the first high-throughput genotyping-by-sequencing in garlic by DArTseq technology. Olive-tree genotyping by HRM analyses proved to be a suitable "closed-tube" approach. In conclusion, both analyses represent a cost-effective methodology for germplasm characterization and genotyping studies. The analyses performed in these chapters could shed some light to help genetic assessment and approaches to study adaptation to fight biotic and abiotic stresses. Which is particularly relevant in the current context of climate change and global warming.

Introducción

Los bancos de germoplasma de plantas son reservorios de biodiversidad. Pueden albergar variedades en múltiples formas, como semillas, arboretos, sembrando cultivos estacionales y manteniendo cultivos *in vitro*. Los registros de los primeros bancos de germoplasma pertenecen a las sociedades egipcia y babilónica. Probablemente, Nikolaj Ivanovič Vavilov fue la primera persona que recalcó la necesidad de crear bancos de germoplasma para el bienestar de la sociedad. Tradicionalmente, las variedades eran almacenadas en función de las características morfológicas. Sin embargo, debido a la plasticidad fenotípica de las plantas, esta clasificación puede dar lugar a sinonimias y homonimias. Esta situación ha desencadenado la necesidad de caracterizar los bancos de germoplasma, no solo por criterios morfológicos, sino también mediante marcadores moleculares. Recientemente, estos últimos han mejorado significativamente. De este modo, han evolucionado desde métodos basados en péptidos a otros basados en ADN. Es más, estos últimos han mejorado enormemente gracias a la reacción en cadena de la polimerasa ("PCR"), incluyendo aproximaciones de alto rendimiento recientes. Esto ha sido posible gracias a la emergencia de tecnologías como las plataformas de secuenciación de segunda generación ("SGS") y de tercera generación ("TGS"). Las aproximaciones utilizadas en la actualidad incluyen las repeticiones de secuencias únicas ("SSR"), los polimorfismos de nucleótidos únicos ("SNP") y el genotipado por secuenciación ("GBS"). En resumen, la caracterización molecular permite una mejor identificación genética, comprensión de las funciones biológicas y búsqueda de relaciones para biología evolutiva. También son utilizadas en biología de la conservación de plantas, estudios de bioseguridad y gestión de bancos de germoplasma. Además, son herramientas excelentes para la mejora asistida, así como para la certificación de la propiedad intelectual y aplicaciones en trazabilidad.

Contenido de la investigación

En esta Tesis Doctoral, bancos de germoplasma de ajo y olivo han sido analizados mediante marcadores moleculares. Brevemente, el ajo ha sido tradicionalmente utilizado en todo el mundo como un ingrediente común en alimentación y como un remedio natural en farmacología y medicina. Esto es debido a sus interesantes atributos beneficiosos como, la reducción de la tensión alta, del colesterol y en la arterioesclerosis. El ajo también tiene efectos preventivos contra el cáncer y actividad antimicrobiana. Por otro lado, la importancia del olivo en innegable. Es uno de las especies más cultivadas en el

mundo, especialmente en la cuenca mediterránea, y es la segunda especie más utilizada en la producción de aceite tras el aceite de palma. El área total de cultivo es diez millones de hectáreas. Sus innumerables propiedades culinarias y medicinales han impulsado su expansión a áreas que no tienen tradición productora ni consumidora.

El principal objetivo de la caracterización del banco de germoplasma de ajo era estudiar la diversidad y estructura genética de 417 muestras del principal bando de germoplasma de ajo localizado en el Instituto Andaluz de Investigación y Formación Agraria, Pesquera, Alimentaria y de la Producción Ecológica (IFAPA) de la Junta de Andalucía, de la Universidad de Córdoba y del Centro de Ensayos de Evaluación de Variedades, localizado en el Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA) en Madrid. La técnica elegida fue DArTseq (tecnología de matriz de diversidad), la cual permite la caracterización en especies sin genoma de referencia u otra información genética previa. De este modo, una colección nuclear fue creada para reducir el número original de entradas del banco, sin perder diversidad genética. Además, las secuencias polimórficas de ajo generadas en el análisis de DArTseq fueron usadas para determinar sus identidades, funciones, términos de Ontología Génica ("GO") y rutas metabólicas mediante la búsqueda de identidades en otras bases de datos de plantas.

Por otro lado, el objetivo principal del capítulo de olivo fue describir un método de "tubo cerrado" rentable para genotipar variedades cuando existe información genética disponible. En este caso, 83 muestras fueron analizadas usado seis marcadores moleculares y el análisis de alta resolución de fusión ("HRM"). Además, aunque el número de marcadores empleado fue bajo, los análisis de caracterización estaban en concordancia con trabajos previos.

Conclusión

Ambas técnicas de marcadores moleculares (DArTseq y HRM) mostraron resultados de genotipado consistentes en función de la información previa de los pasaportes de datos. El tamaño del banco de germoplasma de ajo fue significativamente reducido, lo que indica que en análisis DArTseq es una tecnología adecuada para el genotipado de alto rendimiento sin información genética previa disponible. Hasta nuestro conocimiento, este es el primer genotipado por secuenciación de alto rendimiento en ajo mediante la tecnología DArTseq. El genotipado mediante análisis de HRM en olivo demostró ser una metodología rentable para la caracterización de germoplasma y los

estudios de genotipado. Los análisis realizados en estos capítulos pueden ayudar a clarificar los estudios genéticos y las aproximaciones para estudiar las adaptaciones para hacer frente a estreses bióticos y abióticos. Lo que es particularmente relevante en el contexto actual de cambio climático y calentamiento global.

GENERAL INTRODUCTION

I.1. Plant-germplasm banks

Plant germplasm is a term that comprises any kind of genetic resource from plants or plant structures (such as fruits, seeds or pollen) that is collected or stored for breeding, research, or conservation purposes. When germplasm of different species or varieties is preserved, it is called a germplasm bank. Humans have collected, exchanged, and stored seeds since prehistoric times. The first well-documented records of stored germplasm were found in Egyptian and Babylonian civilizations (Hyland, 1977). There are many kinds of germplasm banks, like traditional ones with seeds, field genebanks with sown plants that are cropped seasonally, arboretums or in vitro cultured, or cryopreserved banks. Likely, Nikolaj Ivanovič Vavilov was the first person that highlighted the importance of storing germplasm as a plant genetic-diversity source. Currently, the N. I. Vavilov Research Institute of Plant Industry (VIR) in Russia is one of the most important germplasm banks. Other relevant institutions are the Consultative Group for International Agricultural Research (CGIAR), the Kew's Millennium Seed Bank Project (England), the Institute of Crop Germplasm Resources of the Chinese Academy of Agricultural Sciences (ICGR-CAAS) in China, the Nordic Genetic Resource Center (NordGen) in Sweden, Svalbard Global Seed Vault in Norway, the National Bureau of Plant Genetic Resources (NBPGR) in India, and the National Center for Genetic Resources Preservation (NCGRP) that belongs to the United States Department of Agriculture (USDA) in USA (Sachs, 2009; FAO, 2010).

There are monospecific banks, focused only in different varieties of the same species, or germplasm banks that harbor different species. Traditionally, germplasm has been stored for breeding purposes, that is, to improve crop plant features or productivity in edible species, or those interesting for feed, forage, turf, and fiber, or other industrial processes. Additional functions were to preserve plants for other uses rather than breeding, like ornamental plants or wild plants that could have medicinal properties (National Research Council Staff, 1990). Nevertheless, the main role of germplasm banks nowadays is to find and preserve germplasm from different varieties or species. Another function of banks is to deliver seeds or other kind of germplasm mainly to researchers or breeders. Currently, the development of new molecular techniques, along with the

increasing number of genebanks, has proved to be useful to enlighten some aspects of crop or plant biology, such as conservation biology, domestication, genetic erosion, or genetic vulnerability in basic and applied research fields (Tanksley and McCouch, 1997; Sachs, 2009).

Stored varieties range from wild varieties, primitive landraces, or adapted ecotypes. They may harbor a great source of genetic variability, to inbred or hybrid lines, superior varieties and elite lines with specific and interesting combinations of characteristics. Many germplasm banks store exotic varieties aside adapted ones. This way, there are more possibilities of response in an event of drastic climate variation. Furthermore, germplasm banks are the centerpiece of food security, nutrition security, and economic development in many countries. Since 19th century, countries started to store germplasm, not only for breeding or researching purposes, but also to ensure food prices to avoid economic crises of sudden and steep price increases, which may lead to hunger and poverty (National Research Council Staff, 1990; FAO, 2010).

Nowadays, the total number of accessions preserved in *ex situ* banks is approximately 7.4 million. Notwithstanding, it has been estimated that only 25-30% of stored accessions are unique, whereas, the remaining 70-75% are duplicates. Many countries have highlighted the necessity of developing better techniques to improve genetic-diversity monitoring, and to establish thresholds and baselines. In fact, public awareness regarding the importance of maintaining genetic diversity is increasing (FAO, 2010).

The main problem in germplasm banks is their conservation and management. To solve this, it is vital to use new biotechnologies like molecular markers or geographic information systems (GIS) to characterize plant diversity. Many collections are exclusively characterized by agromorphological criteria, which can cause homonymy (same name for genetically-different cultivars) or synonymy (same cultivars with different names) (Zhao et al., 2010; Govindaraj et al., 2015). Many genebanks are increasing their collections without being able to evaluate genetic diversity. This is a

major concern as the costs in terms of labor, as well as space and time needed for regeneration and replenishment are high. This may lead to both a higher amount of duplicated accessions and, at the same time, loss of many entries due to the inability to properly manage them (National Research Council Staff, 1990).

Yet, this does not mean that collections must not be duplicated. Storing identical samples in different locations is the only way to lower the risk of losing accessions due to technical errors or unexpected events. Another option is to share accessions among germplasm banks. This way, the same genetic diversity is represented in different collections. Conversely, the current coverage of this representation is uneven, being well covered only for species such as wheat or rice (National Research Council Staff, 1990). Hence, it is necessary to promote, coordinate and facilitate the use of plant-genetic resources between curators and breeders, and and *in situ* and *ex situ* conservation. These collaborations should always be done under policies, regulations, and legislation promoting exchanges within and among countries, which, in many cases, still need to be developed (FAO, 2010).

According to Food and Agriculture Organization of the United Nations (FAO), there are more than 1,750 individual germplasm banks worldwide and 2,500 botanic gardens (FAO, 2010). In Spain, currently there are 71,330 accessions preserved in 33 institutions, being 31,393 local varieties. In short, in seed banks there are approximately 12,500 cereal entries, 15,600 legumes, 18,900 of horticultural crops, 8,200 forage and turf accessions, 1,100 industrial crops, 1000 aromatic and medicinal plants, 7,100 wild varieties, and 150 ornamental species. In arboretums, there are 3,700 fruit trees, 1,600 grape-wine cultivars, 300 olive trees, 576 forestry species, and 170 ornamentals. In total, there are around 1,000 genera and 3,800 different species from 130 countries.

I.2. Molecular markers for genetic characterization

As it has been mentioned in the previous section, characterizing germplasm is vital in order to understand and manage genetic diversity and to reduce costs of maintenance in germplasm banks. Nowadays, the chosen tools for molecular characterization are based on molecular markers. This technology has drastically evolved in the last 60 years, changing from non-DNA-based methods like isozymes to DNA-based ones, increasing complexity and sophistication. As regards DNA-based methods, they have evolved from hybridization-based to PCR-based strategies, improving precision, resolution, and the feasibility of high-throughput techniques. Early methods such as Amplified Fragment Length Polymorphism (AFLP), Restriction Fragment Length Polymorphism (RFLP), or Random Amplified Polymorphic DNA (RAPD) were less accurate and more time consuming than current ones. Even though, they are still useful in many cases. Yet, the improvement of sequencing techniques and the appearance of Second-Generation Sequencing (SGS) and Third-Generation Sequencing (TGS) platforms have allowed high-throughput sequencing. Another advantage of SGS and TGS platforms is the possibility to develop or to improve previous-existing markers such as Simple Sequence Repeat (SSR), Single Nucleotide Polymorphisms (SNP) and, recently, Genotyping-by-Sequencing (GBS) (Henry, 2012; Dorado et al., 2015). These improvements have allowed to better genetic identification, understanding of biological functions, finding of genetic relationships for evolutionary biology, plant conservation biology, biosecurity studies, germplasm-bank management, and breeders applications, as well as intellectual-property rights or traceability applications (Henry, 2012).

Plant domestication started during a historical changing frame from Paleolithic to Neolithic period, approximately 11,000 years ago, probably due to climatic change events that happened during that period. During domestication, men empirically selected wild plants with desirable characteristics for cultivation. Typical traits were those involving better synchronicity of seed germination or harvesting times, better performance against abiotic or biotic stress responses, and better organoleptic characteristics. To do so, seed or other germplasm was kept from season to season, selecting in each one those plants that showed the most favorable characteristics. This way, farmers started to reduce

genetic diversity even without having any knowledge of the bare existence of it (Harlan, 1992).

Nevertheless, classic breeding has limitations. It was not until the "Green Revolution", and afterwards, the emergence of molecular markers, when molecular genetic assays took place. By this time, Mendel's work was rediscovered, shedding light on the mechanisms governing inheritance of phenotypic characteristics. This allowed breeders to combine traditional-breeding techniques with controlled crosses to find more desirable traits, creating new plant varieties with different genetic backgrounds in shorter periods of time. After the introduction of molecular techniques and DNA-based markers, plant-breeding programs have drastically evolved. This is due to four main reasons: i) molecular markers are inheritable, as they are targeted to DNA sequences inherited along varieties or species, increasing predictive power for breeders; ii) genetic information is found in any kind of tissue, so markers may help to predict phenological characteristics, maturity, size or quality traits even from seeds before sow; iii) the numbers of markers that can be developed for genomes is almost unlimited up-to-date; and iv) they can be used to develop Quantitative Trait Loci (QTL) (Henry, 2012).

For a few decades, breeding programs used QTL and Marker-Assisted or Aided Selection (MAS) (Mohan et al., 1997). However, this approach has a major drawback when desirable characteristics are physically distant from the true allele that was marked, as linkage patterns and regions' associations change with increasing distances. Consequently, this technique has a distance limitation. Additionally, other problems are the fact that i) the two parental lines needed for crosses must have big differences in the phenotype of interest; ii) the possible recombination found may be limited when mapping some populations; and iii) alleles of mapped populations are exclusively derived from parental lines, which may require the use of Genome-Wide Association Studies (GWAS). Nevertheless, this has been overcome with SGS and TGS technologies (Varshney et al., 2014). They allow to get large amounts of molecular data that provides useful information to develop markers englobing many traits, helping to develop more QTL and to perform better MAS. Furthermore, such approach allows to use larger panels of unrelated individuals, increasing the possibility of finding new recombination patterns. This

strategy and the improvement of bioinformatic software are allowing the use of GWAS approaches at a high-throughput level. What is more, sequencing methodologies are in continuous development, increasing throughput and reducing time and cost. Testing and implementing these improvements in the breeding field do and will revolutionize breeding programs in terms of time and success (Henry, 2012).

In addition to plant breeding, other fields have also taken advantage of molecular-marker development. For instance, plant conservation, biosecurity, traceability and crop characterization. Particularly, crop characterization has been favored by the betterment of genetic variation detection by PCR-based markers. Genome organization, evolutionary relationships, estimates of Linkage Disequilibrium (LD) or genetic variations are examples of common techniques used routinely in research groups. Within PCR-based assays, those frequently used in plants are mainly SSR, SNP or Expressed-Sequence Tags (EST). Moreover, there are some other techniques, for instance, identification of insertions/deletions (indels), presence/absence of variations (PAV), and copy number variations (CNV). In addition, the emergence of SGS and TGS technologies in the last decade and the current reduction in their costs, is allowing the use of new techniques such as genome-wide SNP discovery in many species. Nevertheless, this is limited to species with sequenced genomes. *De novo* sequencing of plants genomes may be challenging due to i) large genome size; ii) high ploidy iii) high heterozygosity rates, iv) repetition rates, v) high frequency of pseudogenes and transposons; and vi) high copy number of chloroplast and mitochondria organelles. This may increase the costs related to isolation of DNA and bioinformatic analyses, besides increasing turnaround times (Schatz et al., 2012).

As indicated above, SGS and TGS technologies are allowing to increase sequence information and SNP discovery, which can therefore be used for resequencing other sets of individuals. Conversely, this still requires a great endeavor of time and costs. The necessity of overcoming the drawbacks mentioned above and of finding routine techniques for plant genetics, genomics, and molecular breeding has led to the development of GBS protocols (Henry, 2012). Diversity Arrays (DArTseq) technology is a useful GBS approach to study non-model plants where genetic information is scarce.

DArTseq was developed as an extension of previous DArT technology. It performs a sequence of steps in order to reduce genome complexity and to maximize genome polymorphism among samples. First, complexity reduction is done by digesting DNA with a combination of restriction enzymes. Then, polymorphic fragments are selected. Those fragments are cloned into *Escherichia coli* bacteria to create a library. Finally, a hybridization is performed and the signal detected is translated as a presence/absence pattern. In addition to this, for DArTseq markers, after hybridization, the generated library is amplified by Polymerase Chain-Reaction (PCR). Then, amplicons are cleaned, evaluated by capillary electrophoresis sizing and sequenced. The generated sequence file is screened looking for SNP and SilicoDArT markers. The final result is a presence/absence (1 and 0, respectively) matrix (Jaccoud et al., 2001; Gupta et al., 2008; Kilian et al., 2012).

New technologies are currently in development for ultra-high-throughput genotyping platforms. This way, new SGS and TGS-based molecular-marker systems may substitute array-based ones. Promising techniques include: i) Reduced-Representation Sequencing (RRS), which includes Reduced-Representation Libraries (RRL) and Complexity Reduction of Polymorphic Sequences (CRoPS); ii) Restriction-site-Associated DNA sequencing (RAD-seq): iii) Low-Coverage Sequencing for Genotyping, such as GBS (described above); and iv) Multiplexed Shotgun Genotyping (MSG). These technologies are applicable to model and non-model organisms (Davey et al., 2011). Moreover, these novel approaches will revolutionize platforms for plant genotyping and plant breeding, improving already used techniques such as GWAS and Genome-Wide Selection (GWS). Notwithstanding, not every marker system is suitable to address all problems in all species, since each one may have different drawbacks (Henry, 2012).

I.3. Garlic (*Allium sativum*)

Garlic (*Allium sativum* L.) is a bulb vegetable native of Middle Asia but spread for first to China, the Near East, the Mediterranean regions, Central and Southern Europe, Northern Africa (Egypt), Mexico, and then worldwide for agricultural purposes (Maaß and Klaas, 1995; Cardelle-Cobas et al., 2010). Taxonomically, the genus *Allium* belongs to *Liliaceae* family, order *Asparagales*, subclass *Liliidae* (Blanca et al., 2011). By phylogenetic studies it has been seen that *A. longicuspis* seems to be garlic ancestor (Cardelle-Cobas et al., 2010). Since early times, garlic has been a valuable product as a common ingredient in multiple cultures' diet and as a natural remedy. Due to its numerous compounds, it has beneficial effects for high blood pressure, cholesterol, atherosclerosis, has preventive effects to develop some sorts of cancer, and antimicrobial activity (Maaß and Klaas, 1995; Ankri and Mirelman, 1999; Lanzotti, 2006; Pittler and Ernst, 2007; Ma et al., 2009). Nowadays agronomic and industrial interest in garlic joins the culinary and pharmaceutical studies (Kim et al., 2010).

Garlic is known to be a sterile plant that reproduces by vegetative propagation. Conversely, some fertile varieties were found in Central Asia. They are commonly called bolting, and the common and unfertile varieties are called non-bolting. Bolting varieties are able to produce flower-bearing stems and flowers, but seed are not usually fertile (Cholakova, 2000; Shemesh-Mayer et al., 2015; Tchórzewska et al., 2015). In morphological and anatomical studies, interference during teguments formation, and the lack of micropylar channel were observed. On the other hand, non-bolting varieties do not produce flower-bearing stems and reproduce asexually. Some studies have focused in detecting what differentiates bolting and non-bolting varieties and which problems occur during gametophyte generation. In a recent research, where proteomics techniques were used, bolting fertile *Allium* species, *A. tuberosum,* and non-bolting species, *A. sativum* were compared to detect these disorders. In order to do so, Two-Dimensional Gel Electrophoresis (2-DE) protein profiles from *A. sativum* ovules were compared with ovules from *A. tuberosum* to find putative associations in sterility by comparing the differences in protein plot profiles (Winiarczyk and Kosmala, 2009). Homologous studies with male gametogenesis have been done. In this study, sterile *A. sativum* male plants

were compared with *A. leucanthum* and *A. ampeloprasum*. Phenological, morphological, anatomical, pollen viability, stigma receptivity, and phylogenetic analyses were performed by DNA sequencing and 2-DE protein profiles to do fertility comparisons (Shemesh Mayer et al., 2013).

In addition to proteomic assessments, studies characterizing garlic DNA and developing molecular markers for garlic have been published in the last decades. However, compared to other minor crops, the number of microsatellites developed for garlic is still low (Ma et al., 2009; Zhao et al., 2010; Mukherjee and Banerjee, 2013; da Cunha et al., 2014). This is probably due to the fact that variability is reduced because garlic genome is big, there are many duplications, and it reproduces asexually (Ovesná et al., 2014). For instance, (Maaß and Klaas, 1995) used isozymes and RAPD, (Volk et al., 2004) developed AFLP markers, Brazilian and Argentinian garlic varieties has been assessed based on RAPD markers and corroborated with previous morphological and AFLP studies (Lampasona et al., 2003; Buso et al., 2008). (Kim et al., 2009; Bhasi et al., 2010) published databases of EST, and (Ma et al., 2009; Zhao et al., 2010; da Cunha et al., 2014; Ovesná et al., 2014; Ipek et al., 2015) have developed specific SSR makers.

With current genomic tools, the improvement of NGS technologies, and the available sequence data of garlic and related species, the works described in this Doctoral Thesis should shed light on garlic biology and will be useful for future work and for the search of gene sequences with interest in applied fields as breeding in agriculture or pharmacy.

I.4. Olive tree (*Olea europaea*)

Olive tree (*Olea europaea* L.) is diploid and allogamous, with 23 pairs of chromosomes. Such high number of chromosomes and the lack of homology among them indicate that its origin could be allopolyploid, by hybridizations between species near to *Olea* genus (Bracci et al., 2011). This tree is included in *Oleaceae* family, composed by 24 genera and 600 species of evergreen shrubs and trees. *Olea* genus is composed by 33 species, from which olive tree is the only one with edible fruits (Besnard et al., 2009). Such genus is divided in three subgenera: *Paniculatae, Tetrapilus,* and *Olea.* Furthermore, *Olea europaea* has six subspecies, according to morphological characteristics and distribution area: *europaea* in the Mediterranean basin, *cuspidata* in South-Occidental Asia and South-Oriental Africa, *laperrinei* in Saharan region, *maroccana* in Morocco, *cerasiformis* in Madeira Island, and *guanchica* in Canary Islands. *Olea europaea* has two botanic varieties: *sativa*, in which cultivated varieties are included, and *oleaster* or *sylvestris*, which is wild (Green, 2002).

Olive tree is one of the most cultivated species worldwide and it is the second species for oil production after palm oil, having a total cultivated area of ten million hectares (Baldoni and Belaj, 2009; Bracci et al., 2011). The most accepted theory about the origin of olive-tree cultivation was outlined by Vavilov, who stablished its origin in Syria and Iran. From there, it spread towards West and all Mediterranean basin, following commercial routes stablished by Phoenicians, Greeks and Romans (Bartolini et al., 2002). Romans played a main role in olive-tree expansion due to its massive colonization. Afterwards, it expanded to American colonies, Asia and Australia (Fernández Escobar et al., 2008). In Spain, olive tree is the crop that has the highest cultivated surface, 2.5 million hectares, representing 5% of total crop area. Furthermore, in Andalusia, this surface is 1.5 million hectares, which represents 18% of total crop area in this region (Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente, 2015).

Hybridizations, along with allogamy, self-incompatibility, adaptation, and selection by breeders have led to the emergence of many cultivars worldwide. According to Bartolini (2008), there are 1,250 different olive-tree varieties cultivated, on

approximately 50 countries, mainly in Southern Europe. Italy harbors the higher amount of varieties (600), followed by Spain (183), France (88) and Greece (52) (Baldoni and Belaj, 2009). Gene banks, and specially the Worldwide Olive Germplasm Bank of Córdoba (WOGBC) in Spain, have been created to manage this biodiversity. Unfortunately, cultivars with the same genotype showing different agro-morphological characteristics due to phenotypic plasticity, may lead to a case of synonymy. Equally, cultivars with the same phenotype can have different genotypes, which may lead to homonymy.

Therefore, it is vital to genotype olive-tree biodiversity of all cultivars. Up-to-date, there have been many works in this scope, helping to untangle cultivar diversity. They have included AFLP (Angiolillo et al., 1999), RAPD (Besnard et al., 2001) and Sequence Characterized Amplified Region (SCAR) markers (Busconi et al., 2006; Bracci et al., 2011). Subsequent works used SSR and SNP, including also High-Resolution Melting (HRM) Analysis, and, more recently, DArTseq technology has been used (Muleo et al., 2009; Belaj et al., 2012; Domínguez-García et al., 2012; Atienza et al., 2013; Las Casas et al., 2014; Trujillo et al., 2014; Simko, 2016). Moreover, olive-tree transcriptomes and genomes have already been sequenced (Muñoz-Mérida et al., 2013; Cruz et al., 2016).

The third chapter of this thesis applies HRM analysis as an example of a “closed-tube” technique for genotyping germplasm diversity. This can be cost- and time-effective as well as very informative if there is previous genetic information.

I. 5. General objective

The main objective of this work is to analyze garlic and olive-tree germplasm banks, using molecular markers and bioinformatic tools to identify polymorphisms and sequences of interest.

I.6. References

ANGIOLILLO, A., M. MENCUCCINI, and L. BALDONI. 1999. Olive genetic diversity assessed using amplified fragment length polymorphisms. *Theoretical and Applied Genetics* 98: 411–421.

ANKRI, S., and D. MIRELMAN. 1999. Antimicrobial properties of allicin from garlic. *Microbes and Infection* 1: 125–129.

ATIENZA, S.G., R. DE LA ROSA, M.C. DOMÍNGUEZ-GARCÍA, A. MARTÍN, A. KILIAN, and A. BELAJ. 2013. Use of DArT markers as a means of better management of the diversity of olive cultivars. *Food Research International* 54: 2045–2053.

BALDONI, L., and A. BELAJ. 2009. Olive. *In* J. Vollmann, and I. Rajcan [eds.], Oil crops, Handbook of Plant Breeding, 397–421. Springer New York.

BARTOLINI, G. 2008. Olive Germplasm (*Olea europaea* L.).

BARTOLINI, G., R. PETRUCCELLI, and F. AND A.O. OF THE U. NATIONS. 2002. Classification, origin, diffusion and history of the Olive. *Food and Agriculture Organization* (FAO).

BELAJ, A., M. DEL C. DOMINGUEZ-GARCÍA, S.G. ATIENZA, N.M. URDÍROZ, R.D. LA ROSA, Z. SATOVIC, A. MARTÍN, A. KILIAN, I. TRUJILLO, V. VALPUESTA, and C. DEL RÍO. 2012. Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genetics & Genomes* 8: 365–378.

BESNARD, G., P. BARADAT, and A. BERVILLÉ. 2001. Genetic relationships in the olive (*Olea europaea* L.) reflect multilocal selection of cultivars. *Theoretical and Applied Genetics* 102: 251–258.

BESNARD, G., R. RUBIO DE CASAS, P.-A. CHRISTIN, and P. VARGAS. 2009. Phylogenetics of *Olea* (*Oleaceae*) based on plastid and nuclear ribosomal DNA sequences:

Tertiary climatic shifts and lineage differentiation times. *Annals of Botany* 104: 143–160.

BHASI, A., D. SENALIK, P.W. SIMON, B. KUMAR, V. MANIKANDAN, P. PHILIP, and P. SENAPATHY. 2010. RoBuST: an integrated genomics resource for the root and bulb crop families *Apiaceae* and *Alliaceae*. *BMC Plant Biology* 10: 161.

BLANCA, G., B. CABEZUDO, M. CUETO, C. SALAZAR, and C. MORALES TORRES. 2011. Flora vascular de Andalucía Oriental. 2ª edición. Universidades de Almería, Granada, Jaén y Málaga, Granada.

BRACCI, T., M. BUSCONI, C. FOGHER, and L. SEBASTIANI. 2011. Molecular studies in olive (*Olea europaea* L.): overview on DNA markers applications and recent advances in genome analysis. *Plant Cell Reports* 30: 449–462.

BUSCONI, M., L. SEBASTIANI, and C. FOGHER. 2006. Development of SCAR markers for germplasm characterisation in olive tree (*Olea europea* L.). *Molecular Breeding* 17: 59–68.

BUSO, G.S.C., M.R. PAIVA, A.C. TORRES, F.V. RESENDE, M.A. FERREIRA, J.A. BUSO, and A.N. DUSI. 2008. Genetic diversity studies of Brazilian garlic cultivars and quality control of garlic-clover production. *Genetics and molecular research 7*: 534–541.

CARDELLE-COBAS, A., SORIA, A. C., CORZO-MARTINEZ, M., AND VILLAMIEL, M. 2010. "A comprehensive survey of garlic functionality," *In* Garlic Consumption and Health. M. Pacurar and G. Krejci [eds]: 1–60. Hauppauge: Nova Science Publishers, Inc.

CHOLAKOVA, N. 2000. Application of esterase isozymes for garlic ecotype identification. *Biologia Plantarum* 43: 445–446.

CRUZ, F., I. JULCA, J. GÓMEZ-GARRIDO, D. LOSKA, M. MARCET-HOUBEN, E. CANO, B. GALÁN, L. FRIAS, P. RIBECA, S. DERDAK, M. GUT, M. SÁNCHEZ-FERNÁNDEZ, J. L. GARCÍA, I. G. GUT, P. VARGAS, T. S. ALIOTO and T. GABALDÓN. 2016. Genome sequence of the olive tree, *Olea europaea*. *GigaScience* 5: 29.

DA CUNHA, C.P., F.V. RESENDE, M.I. ZUCCHI, and J.B. PINHEIRO. 2014. SSR-based genetic diversity and structure of garlic accessions from Brazil. *Genetica* 142: 419–431.

DAVEY, J.W., P.A. HOHENLOHE, P.D. ETTER, J.Q. BOONE, J.M. CATCHEN, and M.L. BLAXTER. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12: 499–510.

DOMÍNGUEZ-GARCÍA, M.C., A. BELAJ, R. DE LA ROSA, Z. SATOVIC, K. HELLER-USZYNSKA, A. KILIAN, A. MARTÍN, and S.G. ATIENZA. 2012. Development of DArT markers in olive (*Olea europaea* L.) and usefulness in variability studies and genome mapping. *Scientia Horticulturae* 136: 50–60.

DORADO, G., T. UNVER, H. BUDAK, and P. HERNÁNDEZ. 2015. Molecular markers. *In* Caplan M [ed]. Reference Module in Biomedical Sciences. Biochemistry, Cell Biology and Molecular Biology. Elsevier Amsterdam.

FAO. 2010. The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. Rome.

FERNÁNDEZ ESCOBAR, R., and L. RALLO ROMERO. 2008. Variedades y patrones. *In* E-libro, Corp [eds.] El cultivo del olivo, 37–62. Mundi-Prensa, Madrid.

GOVINDARAJ, M., M. VETRIVENTHAN, and M. SRINIVASAN. 2015. Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genetics research international* 2015: 431487–431487.

GREEN, P.S. 2002. A Revision of *Olea* L. (*Oleaceae*). *Kew Bulletin* 57: 91–140.

GUPTA, P.K., S. RUSTGI, and R.R. MIR. 2008. Array-based high-throughput DNA markers for crop improvement. *Heredity* 101: 5–18.

HARLAN, J. 1992. Crops and man. American Society of Agronomy, Madison, Wisconsin, USA.

HENRY, R.J. 2012. Evolution of DNA marker technology in plants. *In* R. J. Henry [ed.], Molecular markers in plants, 1–19. Blackwell Publishing Ltd.

HYLAND, H.L. 1977. History of U.S. Plant introduction. *Environmental History Review* 2: 26–32.

IPEK, M., N. SAHIN, A. IPEK, A. CANSEV, and P. SIMON. 2015. Development and validation of new SSR markers from expressed regions in the garlic genome. *Scientia Agricola* 72: 41–46.

JACCOUD, D., K. PENG, D. FEINSTEIN, and A. KILIAN. 2001. Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Research* 29: E25.

KILIAN, A., P. WENZL, E. HUTTNER, J. CARLING, L. XIA, H. BLOIS, V. CAIG, K. HELLER-USZYNSKA, D. JACCOUD, C. HOPPER, M. ASCHENBRENNER-KILIAN, M. EVERS, K. PENG, C. CAYLA, P. HOK, G. USZYNSKI. 2012. Diversity arrays technology: a generic genome profiling technology on open platforms. *Methods in molecular biology (Clifton, N.J.)* 888: 67–89.

KIM, A., R. KIM, D. KIM, S. CHOI, A. KANG, S. NAM, and H. PARK. 2010. Identification of a novel garlic cellulase gene. *In* Plant molecular biology reporter, 388–93.

KIM, D.-W., T.-S. JUNG, S.-H. NAM, H.-R. KWON, A. KIM, S.-H. CHAE, S.-H. CHOI, D-W. KIM, R. N. KIM, and H-S. PARK. 2009. GarlicESTdb: an online database and mining tool for garlic EST sequences. *BMC Plant Biology* 9: 61.

LAMPASONA, S.G., L. MARTÍNEZ, and J.L. BURBA. 2003. Genetic diversity among selected Argentinean garlic clones (*Allium sativum* L.) using AFLP (Amplified Fragment Length Polymorphism). *Euphytica* 132: 115–119.

LANZOTTI, V. 2006. The analysis of onion and garlic. *Journal of Chromatography. A* 1112: 3–22.

LAS CASAS, G., F. SCOLLO, G. DISTEFANO, A. CONTINELLA, A. GENTILE, and S. LA MALFA. 2014. Molecular characterization of olive (*Olea europaea* L.) Sicilian cultivars using SSR markers. *Biochemical Systematics and Ecology* 57: 15–19.

MA, K.-H., J.-G. KWAG, W. ZHAO, A. DIXIT, G.-A. LEE, H.-H. KIM, I.-M. CHUNG, N-S KIM, J-S. LEE, J-J. JI, T-S. KIM, and Y-J. PARK. 2009. Isolation and characteristics

of eight novel polymorphic microsatellite loci from the genome of garlic (*Allium sativum* L.). *Scientia Horticulturae* 122: 355–361.

MAAß, H.I., and M. KLAAS. 1995. Infraspecific differentiation of garlic (*Allium sativum* L.) by isozyme and RAPD markers. *Theoretical and applied genetics. Theoretische und angewandte Genetik* 91: 89–97.

MINISTERIO DE AGRICULTURA Y PESCA, ALIMENTACIÓN Y MEDIO AMBIENTE. 2015. Encuesta sobre Superficies y Rendimientos Cultivos (ESYRCE). Available at: http://www.mapama.gob.es/es/estadistica/temas/estadisticas-agrarias/agricultura/esyrce/# [Accessed February 11, 2017].

MOHAN, M., S. NAIR, A. BHAGWAT, T.G. KRISHNA, M. YANO, C.R. BHATIA, and T. SASAKI. 1997. Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding* 3: 87–103.

MUKHERJEE, D., and S. BANERJEE. 2013. Learning and memory promoting effects of crude garlic extract. *Indian Journal of Experimental Biology* 51: 1094–1100.

MULEO, R., M.C. COLAO, D. MIANO, M. CIRILLI, M.C. INTRIERI, L. BALDONI, and E. RUGINI. 2009. Mutation scanning and genotyping by high-resolution DNA melting analysis in olive germplasm. *Genome* 52: 252–260.

MUÑOZ-MÉRIDA, A., J.J. GONZÁLEZ-PLAZA, A. CAÑADA, A.M. BLANCO, M. DEL C. GARCÍA-LÓPEZ, J.M. RODRÍGUEZ, L. PEDROLA, M. D. SICARDO, M. L. HERNÁNDEZ, R. DE LA ROSA, A. BELAJ, M. GIL-BORJA, F. LUQUE, J. M. MARTÍNEZ-RIVAS, D. G. PISANO, O. TRELLES, V. VALPUESTA, and C. R. BEUZÓN. 2013. *De novo* assembly and functional annotation of the Olive (*Olea europaea*) transcriptome. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 20: 93–108.

NATIONAL RESEARCH COUNCIL STAFF. 1990. U.S. National Plant Germplasm System. National Academies Press, Washington, US. Available at: http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10056767 [Accessed April 5, 2017].

OVESNÁ, J., L. LEIŠOVÁ-SVOBODOVÁ, and L. KUČERA. 2014. Microsatellite analysis indicates the specific genetic basis of Czech bolting garlic. *Czech Journal of Genetics and Plant Breeding* 50: 226–234.

PITTLER, M.H., and E. ERNST. 2007. Clinical effectiveness of garlic (*Allium sativum*). *Molecular Nutrition & Food Research* 51: 1382–1385.

SACHS, M.M. 2009. Cereal germplasm resources. *Plant Physiology* 149: 148–151.

SCHATZ, M.C., J. WITKOWSKI, and W.R. MCCOMBIE. 2012. Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biology* 13: 243.

SHEMESH-MAYER, E., K. WINIARCZYK, L. BŁASZCZYK, A. KOSMALA, H.D. RABINOWITCH, and R. KAMENETSKY. 2013. Male gametogenesis and sterility in garlic (*Allium sativum* L.): barriers on the way to fertilization and seed production. *Planta* 237: 103–120.

SHEMESH-MAYER, E., T. BEN-MICHAEL, N. ROTEM, H.D. RABINOWITCH, A. DORON-FAIGENBOIM, A. KOSMALA, D. PERLIKOWSKI, A. SHERMAN, and R. KAMENETSKY. 2015. Garlic (*Allium sativum* L.) fertility: transcriptome and proteome analyses provide insight into flower and pollen development. *Frontiers in Plant Science* 6: 271.

SIMKO, I. 2016. High-resolution DNA melting analysis in plant research. *Trends in Plant Science* 21: 528–537.

TANKSLEY, S.D., and S.R. MCCOUCH. 1997. Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science* 277: 1063–1066.

TCHÓRZEWSKA, D., K. DERYŁO, L. BŁASZCZYK, and K. WINIARCZYK. 2015. Tubulin cytoskeleton during microsporogenesis in the male-sterile genotype of *Allium sativum* L. and fertile *Allium ampeloprasum* L. *Plant Reproduction* 28: 171–182.

TRUJILLO, I., M.A. OJEDA, N.M. URDIROZ, D. POTTER, D. BARRANCO, L. RALLO, and C.M. DIEZ. 2014. Identification of the Worldwide Olive Germplasm Bank of Córdoba (Spain) using SSR and morphological markers. *Tree Genetics & Genomes* 10: 141–155.

VARSHNEY, R.K., R. TERAUCHI, and S.R. MCCOUCH. 2014. Harvesting the promising fruits of genomics: Applying genome sequencing technologies to crop breeding. *PLoS Biology* 12: e1001883.

VOLK, G.M., A.D. HENK, and C.M. RICHARDS. 2004. Genetic diversity among U.S. garlic clones as detected using AFLP methods. *Journal of the American Society for Horticultural Science* 129: 559–569.

WINIARCZYK, K., and A. KOSMALA. 2009. Development of the female gametophyte in the sterile ecotype of the bolting *Allium sativum* L. *Scientia Horticulturae* 121: 353–360.

ZHAO, W., J. CHUNG, G. LEE, K. MA, H. KIM, K. KIM, I. CHUNG, J. K. LEE, N. S. KIM, S. M. KIM, and Y. J. PARK. 2010. Molecular genetic diversity and population structure of a selected core set in garlic and its relatives using novel SSR markers. *Plant Breeding* 130: 46–54.

CHAPTER 1. Assessment of genetic diversity and structure of large garlic (Allium sativum) germplasm bank, by Diversity Arrays Technology "genotyping-by-sequencing" platform (DArTseq)

(Published as EGEA L A., R. MÉRIDA-GARCÍA, A. KILIAN, P. HERNANDEZ, G. DORADO. 2017. Assessment of Genetic Diversity and Structure of Large Garlic (*Allium sativum*) Germplasm Bank, by Diversity Arrays Technology "Genotyping-by-Sequencing" Platform (DArTseq). *Frontiers in Genetics* 8: 98)

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 July 2017
doi: 10.3389/fgene.2017.00098

Assessment of Genetic Diversity and Structure of Large Garlic (*Allium sativum*) Germplasm Bank, by Diversity Arrays Technology "Genotyping-by-Sequencing" Platform (DArTseq)

*Leticia A. Egea[1,2], Rosa Mérida-García[2], Andrzej Kilian[3], Pilar Hernandez[2] and Gabriel Dorado[1]**

[1] Departamento de Bioquímica y Biología Molecular, Campus Rabanales (C6-1-E17), Campus de Excelencia Internacional Agroalimentario (ceiA3), Universidad de Córdoba, Córdoba, Spain, [2] Instituto de Agricultura Sostenible (IAS-CSIC), Campus Alameda del Obispo, Córdoba, Spain, [3] Diversity Arrays Technology Pty. Ltd., Canberra, ACT, Australia

OPEN ACCESS

Edited by:
Samuel A. Cushman, United States Forest Service Rocky Mountain Research Station, United States

Reviewed by:
Turgay Unver, iBG-Izmir, International Biomedicine and Genome Institute, Turkey
Hikmet Budak, Montana State University, United States
Guillaume Besnard, UMR5174 Evolution Et Diversite Biologique (EDB), France

***Correspondence:**
Gabriel Dorado
bb1dopeg@uco.es

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *05 April 2017*
Accepted: *30 June 2017*
Published: *20 July 2017*

Citation:
Egea LA, Mérida-García R, Kilian A, Hernandez P and Dorado G (2017) Assessment of Genetic Diversity and Structure of Large Garlic (Allium sativum) Germplasm Bank, by Diversity Arrays Technology "Genotyping-by-Sequencing" Platform (DArTseq). Front. Genet. 8:98. doi: 10.3389/fgene.2017.00098

Garlic (*Allium sativum*) is used worldwide in cooking and industry, including pharmacology/medicine and cosmetics, for its interesting properties. Identifying redundancies in germplasm blanks to generate core collections is a major concern, mostly in large stocks, in order to reduce space and maintenance costs. Yet, similar appearance and phenotypic plasticity of garlic varieties hinder their morphological classification. Molecular studies are challenging, due to the large and expected complex genome of this species, with asexual reproduction. Classical molecular markers, like isozymes, RAPD, SSR, or AFLP, are not convenient to generate germplasm core-collections for this species. The recent emergence of high-throughput genotyping-by-sequencing (GBS) approaches, like DArTseq, allow to overcome such limitations to characterize and protect genetic diversity. Therefore, such technology was used in this work to: (i) assess genetic diversity and structure of a large garlic-germplasm bank (417 accessions); (ii) create a core collection; (iii) relate genotype to agronomical features; and (iv) describe a cost-effective method to manage genetic diversity in garlic-germplasm banks. Hierarchical-cluster analysis, principal-coordinates analysis and STRUCTURE showed general consistency, generating three main garlic-groups, mostly determined by variety and geographical origin. In addition, high-resolution genotyping identified 286 unique and 131 redundant accessions, used to select a reduced size germplasm-bank core collection. This demonstrates that DArTseq is a cost-effective method to analyze species with large and expected complex genomes, like garlic. To the best of our knowledge, this is the first report of high-throughput genotyping of a large garlic germplasm. This is particularly interesting for garlic adaptation and improvement, to fight biotic and abiotic stresses, in the current context of climate change and global warming.

Keywords: DNA fingerprinting, breeding, phenotype, somatic mutation, second-generation sequencing (SGS), third-generation sequencing (TGS), next-generation sequencing (NGS)

INTRODUCTION

Garlic (*Allium sativum*) is a plant producing an edible bulb, made of storage leaves known as cloves. It is of Asian origin, being *Allium longicuspis* considered its wild ancestor. It belongs to genus *Allium*, which includes almost 1,000 species, such as chive (*Allium schoenoprasum*), leek (*Allium ampeloprasum*), onion and shallot (*Allium cepa*) (Maab and Klaas, 1995; Kamenetsky et al., 2004; Meredith, 2008; Cardelle-Cobas et al., 2010; Pacurar and Krejci, 2010). Garlic has a large diploid genome ($2n = 2x = 16$), of an estimated haploid (1C) size of 15.9 gigabase pairs (Gbp); that is, 32 times larger than rice (*Oryza sativa*). Garlic is sterile (does not produce fertile botanical seeds by sexual reproduction), asexually propagating by its cloves, despite some progress in recent years to restore garlic fertility (Shemesh-Mayer et al., 2015). Besides, cloves must be reproduced every year, since they cannot be stored for longer periods and then germinated, as happens with standard botanical seeds. Such peculiarity adds extra cost and inconvenience to its maintenance, mainly for large germplasm collections. The peculiar garlic reproduction could lead to low genome diversity, since meiosis is not involved in its clonal reproduction by vegetative propagation (Kamenetsky et al., 2015). Yet, garlic shows a surprisingly high biodiversity, as well as environmental-adaptation capacity and phenotypic plasticity (Volk et al., 2004). All that leads to the large number of garlic varieties or cultivars available (traditionally classified by agromorphological characteristics). The reason for that is not fully understood, suggesting a complex genome (Green, 2001), due to its extremely large size containing many multicopy genes and other duplications, including non-coding sequences and tandem repeats (Arumuganathan and Earle, 1991; Jones et al., 2004; Ovesna et al., 2015), which should be better understood once sequenced. So far, partial and total genome duplications have been described (Supplementary Table S1). Additionally, somatic mutations have been also reported for this species, as well as somaclonal variation, differential gene-expression and alternative splicing (Al-Zahim et al., 1999; Rotem et al., 2007; Kamenetsky et al., 2015; Shemesh-Mayer et al., 2015). Probably, transposable elements are also involved in the evolution of this species.

Besides being appreciated in cooking as common seasoning for thousands of years (Cardelle-Cobas et al., 2010), garlic is also used in pharmacology and cosmetics. Indeed, it is known to have medical properties, protecting against different diseases, like, for instance, hypercholesterolemia, hypertension, atherosclerosis, and thrombosis, reducing the risk of developing cardiovascular disease (CVD). Other recognized bioactivities are antimicrobial (albeit being probiotic), antiasthmatic, antioxidant, anticarcinogenic, etc. (Corzo-Martínez et al., 2007; Pacurar and Krejci, 2010; Rana et al., 2011). Indeed, garlic contains bioactive compounds, including, among others: (i) lectins, which have wide applications in biomedicine and biotechnology (Smeets et al., 1997); (ii) peptides with angiotensin I-converting enzyme (ACE) inhibitory activity, being related to its antihypertensive activity (Suetsuna, 1998); and (iii) *N*-feruloyltyramine, which protects against CVD by suppressing platelet activation (Park, 2009). Besides, this species is rich in enzymes with industrial interest; for instance: (i) nucleases (DNase and RNase), with application in molecular biology (Carlsson and Frick, 1964); (ii) cellulases for biotechnological applications, like conversion of biomass into biofuel (Kim et al., 2010); (iii) superoxide dismutases (SOD), which represent a main defense against oxidative stress, being widely used in pharmacology/medicine, cosmetics, food, agriculture, and chemical industries (He et al., 2008; Liu et al., 2011); (iv) proteases/hemagglutinases, with application in medical tests (Parisi et al., 2008); and (v) alliinases (also known as alliinases), that catalyze conversion of alliin to allicin, which is the main therapeutic agent of garlic (Corzo-Martínez et al., 2007; Kim et al., 2010; Rathnasamy et al., 2014).

On the other hand, agricultural practices usually involve cultivation of a reduced number of species and varieties, which may lead to genetic erosion. That is especially relevant for monocultures, which on the other hand are required to feed an exponentially growing human population. It is therefore important to maintain germplasm banks as reservoirs of genetic variability for crop breeding. Thus, such collections may harbor genetic potential to improve productivity and adaptation/resistance to abiotic (drought, salinity, etc.) and biotic (diseases and plagues) stresses (Tanksley and McCouch, 1997). That is particularly relevant in the current frame of climatic change and global warming. Understanding this potential is critical for identification of biodiversity in biological resources and its efficient management, including conservation and selection of genetically divergent accessions to optimize breeding programs (Olukolu et al., 2012).

Yet, germplasm banks may be generated as mere raw collections of varieties over many years, being classified by criteria based on phenotypic/agronomic traits (passport data). That could lead to both homonymy (same name for genetically different cultivars) and duplications or synonymy (same cultivars with different names). That is especially problematic for species with similar appearance and significant phenotypic plasticity, like garlic. Thus, efficient identification of biodiversity is of paramount importance to manage and maintain such genetic-resources (Govindaraj et al., 2015). That is relevant not only to identify genuine variability for breeding purposes, but also to reduce space and maintenance costs, especially for large germplasm banks, generating reduced, albeit representative, core collections (Zhao et al., 2010).

The role of molecular markers as a tool for genetic analyses and crop improvement has gained importance through the years, as we have reviewed (Dorado et al., 2015c). Their use has become common in model species and important crops. Indeed, genetic diversity and polymorphism assessments are major priorities in plant and crop-breeding studies (Nybom and Bartish, 2000). Large-scale identification of molecular markers like single-nucleotide polymorphism (SNP) on genome and transcriptome represent interesting approaches (Ipek et al., 2016; Akpinar et al., 2017). Classical molecular-markers to assess genetic diversity and polymorphism in garlic have been described (Ovesná et al., 2014; Ipek et al., 2015). Among others, they include isozymes, random-amplified polymorphic DNA (RAPD) (Maab and Klaas, 1995), simple-sequence repeats (SSR) (DaCunha et al., 2014), amplified-fragment length polymorphism (AFLP) (Ipek et al.,

2005) and insertions-deletions (InDel) (Wang et al., 2016). Yet, such analyses of genetic diversity in this species are challenging (Kim et al., 2009).

Fortunately, recent technological developments overcome previous limitations. They include second-generation sequencing (SGS) and third-generation sequencing (TGS) approaches, sometimes known by the ambiguous next-generation sequencing (NGS) terminology, as we have reviewed (Dorado et al., 2015b). Thus, a high-throughput genotyping-by-sequencing (GBS) technology (DArTseq) has been developed. It combines diversity arrays technology (DArT) complexity reduction methods with SGS/TGS (Kilian et al., 2012; Courtois et al., 2013; Cruz et al., 2013; Raman et al., 2014), allowing to identify SNP. DArT markers are polymorphic segments of DNA that are found at specific genome sites, after complexity reduction, being detected by hybridization. Those markers may show dominant or codominant inheritance (Gupta et al., 2008). DArT markers exploit DNA-microarray platforms to analyze DNA polymorphisms, without requiring previous DNA-sequence knowledge. Their applications include genetic fingerprinting, like whole-genome profiling for molecular breeding, germplasm characterization and genetic mapping, among others (Jaccoud et al., 2001). DArTseq can be optimized for each organism and application, by selecting the most appropriate complexity-reduction method (both size of representation and fraction of selected genome for assays). This is particularly relevant for garlic, which has a large and expected complex genome, as previously described. Therefore, DArTseq has been used in the present work as a proof-of-concept, to analyze a large garlic-germplasm bank.

The main goals of this study are: (i) assess genetic diversity and structure of a large garlic-germplasm bank; (ii) create a core collection to reduce the number of original accessions, without losing genetic diversity; (iii) relate genotype to agronomical features; and (iv) describe a cost-effective method to manage genetic diversity that could be applied to germplasm banks and breeding projects of garlic and other species.

MATERIALS AND METHODS

Plant Material and DNA Isolation

A total of 417 *a priori* different garlic entries collected in Spain (some of them being originally derived from other countries) were used for DArTseq analyses: 408 from the main Garlic-Germplasm Bank at "Instituto Andaluz de Investigación y Formación Agraria, Pesquera, Alimentaria y de la Producción Ecológica" (IFAPA) of "Junta de Andalucía" in Cordoba; five from Cordoba University (C1 to C5); and four (G, K, L, and M) from "Centro de Ensayos de Evaluación de Variedades" at "Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria" (INIA) in Madrid (Supplementary Table S1). Garlic leaves were frozen in liquid nitrogen and stored at −80°C until needed.

DNA was isolated using cetyl trimethylammonium bromide (CTAB) protocol (Murray and Thompson, 1980), as we have optimized (Hernandez et al., 2001). It was dissolved in Tris-Na_2EDTA (TE; pH 8) and stored at 4°C. Isolated DNA was quantified by NanoDrop 2000c (Thermo Fisher Scientific, Waltham, MA, United States) and segregated by 1% (w/v) agarose [from United States Biological (Salem, MA, United States)] gel electrophoresis (AGE). Then it was stained with ethidium bromide from Sigma–Aldrich (St. Louis, MO, United States). Resulting DNA was visualized under ultraviolet (UV) light for quality evaluation, using a Molecular Imager VersaDoc MP 4000 System from Bio-Rad (Hercules, CA, United States). Additionally, DNA digestions with the frequent-cutter *Tru*1I restriction enzyme (RE; cutting at T$^{|}$ TA$_{|}$ A) from Thermo Fisher Scientific were performed, in order to check DNA quality and absence of contaminating nucleases.

DArTseq

DArTseq method from Diversity Arrays Technology (Canberra, ACT, Australia) is described elsewhere[1]. In short, the following steps were carried out: (i) complexity reduction, in which genomic DNA was digested with a combination of restriction enzymes. Then, adapters were ligated and only polymorphic fragments were selected. In this way, this technique allowed to exclusively focus in those sections of the genome which are interesting for genetic-diversity analyses, due to their polymorphism; (ii) polymorphic fragments were cloned into *Escherichia coli* bacteria to create a library. Each *E. coli* colony should contain one of those fragments; (iii) the generated library was amplified by polymerase chain-reaction (PCR), as we have reviewed (Dorado et al., 2015a); (iv) amplicons were cleaned and evaluated by capillary electrophoresis sizing; (v) fragments were sequenced; (vi) A FASTQ file was created with generated sequencing reads, including sequences from 30 to 60 base pairs (bp) of polymorphic fragments; (vii) an internal alignment was performed, using other reads from the library (this step is carried out in case of incomplete or absent reference genome, like in the present work); (viii) SNP and SilicoDArT markers were searched and filtered using algorithms; and (ix) resulting data were two presence/absence (1 and 0, respectively) matrices. One contained SNP and the other SilicoDArT markers, where each column represented an individual and each row a marker (Kilian et al., 2012).

In our case, four methods of complexity reduction were tested in garlic (data not shown), selecting the *Pst*I-*Nsp*I restriction enzymes (cutting at G| TGCA| G and R| CATG| Y, respectively). Briefly, DNA samples were processed in digestion/ligation reactions as previously described (Kilian et al., 2012), but replacing a single *Pst*I-compatible adaptor with two different adaptors, corresponding to two different RE overhangs. The *Pst*I-compatible adapter was designed to include flowcell-attachment sequence from Illumina (San Diego, CA, United States), sequencing-primer sequence and "staggered" barcode (varying-length region), similar to previously reported (Elshire et al., 2011). Reverse adapter contained flowcell-attachment region and *Nsp*I-compatible overhang sequence. Interestingly, an overrepresented sequence from cytoplasmic (chloroplastic) DNA, corresponding to >10% of total sequences, was identified (after initial optimization) in many *Pst*I-*Nsp*I garlic-library

[1]http://www.diversityarrays.com/dart-application-dartseq

samples. A cut site for *Alw*I (cutting at GGATCNNNN| N|) was identified within this overrepresented sequence, and thus such restriction enzyme was included in the digestion-ligation step of library construction. Only "mixed fragments" (*Pst*I-*Nsp*I) which did not have *Alw*I site were effectively amplified in 30 rounds of PCR, using the following reaction profile: (i) denaturation at 94°C for 1 min; (ii) 30 cycles [94°C for 20 s (denaturation), 58°C for 30 s (primer annealing) and 72°C for 45 s (primer extension)]; and (iii) final polymerization at 72°C for 7 min. Equimolar amounts of PCR amplicons from each sample reaction of 96-well microtiter plates were bulked and applied to c-Bot (Illumina) bridge PCR, followed by sequencing on HiSeq 2000 sequencing system from the same manufacturer. Single-read sequencing reactions were run for 77 cycles.

Sequences generated from each lane were processed using DArT analytical-pipelines. In the primary one, Fast-Alignment Sequence Tools Q (FASTQ) files were first processed. Thus, poor-quality sequences were filtered-away, applying more stringent selection criteria to the barcode region, as compared to the rest of the sequence. Assignments of sequences to specific samples in the "barcode split" step were very reliable. This way, approximately 2,000,000 sequences per barcode/sample were identified and used in marker calling. Finally, identical sequences were collapsed into "fastqcoll" files. These were "groomed" using the DArT PL's C++ algorithm, which corrects low-quality bases from singleton-tags into correct bases, using collapsed tags with multiple members as template.

Groomed fastqcoll files were used in the secondary pipeline (presence/absence of restriction fragments in representation), by DArT, PL, SNP, and SilicoDArT calling algorithms (DArTsoft version 14). In total, 33,423 presence/absence markers were generated. All tags from all libraries included in the DArTsoft analyses were clustered using the DArT PL's C++ algorithm (threshold distance of 3), for SNP calling. That was followed by cluster parsing into separate SNP loci, using a range of technical parameters; especially the balance of read counts for allelic pairs. Additional selection criteria were added to the algorithm, based on previous experience with analyses of approximately 1,000-controlled cross populations (data not shown). Testing for Mendelian distribution of alleles in these previous populations facilitated selection of technical parameters, discriminating well-true allelic variants from paralogous sequences. In addition, multiple samples were processed from DNA to allelic calls, as technical replicates and scoring consistency was used as the main selection criteria for high-quality/low error-rate markers. Calling quality was assured by high average-read-depth per locus (average across all markers was over 10 reads/locus).

Genetic Diversity and Structure Assessments

Three different analyses were performed, in order to study genetic diversity and structure of germplasm-bank accessions. After creating the SNP and SilicoDArT marker scoring matrices, a Gower's distance matrix was generated. Gower's distance is a coefficient that measures similarity between two samples, based on logical (absence/presence) information differing for several variables (Gower, 1971). These data were used to determine genetically redundant samples. Secondly, a hierarchical cluster-analysis was done with the "pvclust" R package (Suzuki and Shimodaira, 2015). The phylogenetic tree (dendrogram) was computed with a complete-linkage method. By doing complete-linkage clustering (agglomerative hierarchical clustering method), each element of a distance matrix was first individually clustered. Then, each sample was combined into a new cluster, according to the shortest distance (Defays, 1977). Besides previous tests, a principal-coordinates analysis (PCoA; also known as classical multidimensional scaling, Torgerson Scaling or Torgerson-Gower scaling) was also carried out, using R software version 3.2.2 (R-Development-Core-Team, 2015). Additionally, STRUCTURE software version 2.3.4 (Pritchard et al., 2000) was used to study genetic structure. The chosen parameters were five iterations, K ranging from 1 to 3, with a burnin length of 10,000 and 20,000 Markov Chain Monte Carlo (MCMC) repetitions after burnin.

RESULTS

DArTseq Analyses

A total of 417 garlic samples were analyzed using SilicoDArT markers (representing presence/absence of restriction fragments in DArT genomic representations) and SNP data. A total of 14,392 SNP were used for the analyses. DArTseq markers allowed identifying 286 unique (Supplementary Table S2) and 131 redundant samples. The latter were divided into 19 groups, showing a variable amount of individuals (two to 53; Supplementary Table S3). For instance, in group 1, samples 717 and 718 were from the same province (Jaen, Spain). Spanish White varieties were mainly associated in groups 2 and 3 (samples 238, 452, and 461, all from northern Spain). Additionally, for group 2, there was an internal structure between regions. Samples 335, 424, 433, 434, 457, 464, and 467 were from northern Spanish provinces; samples 360 and 368 came from Caceres (Spain) and samples 127, 130, and 553 from southern Spanish provinces. Groups 4 and 7 to 10 included Spanish Purple varieties. Particularly, samples in group 4 were all from Castilla-Leon (Spain). Group 7 was the most numerous, with a total amount of 53 redundant samples. Interestingly, some associations by province were found in this group. Thus, samples 2, 59, 486, and 489 were all from northern regions; samples 21, 37, and 366 from central provinces; and samples 3, 85, 107, 110, 125, 131, 139, 150, 171, 225, 344, 356, 715, and 720 were from southern provinces. Two samples (14 and 280) from Taiwan, were also included in group 7. On the other hand, no associations were found for groups 5, 6, and 11 to 19.

Germplasm-Diversity Assessments

The 417 garlic samples were further analyzed, in order to assess their genetic diversity and structure, to eliminate redundant accessions, and thus generate the germplasm-bank core collection. Two different analyses were performed: hierarchical cluster computed by complete-linkage method and PCoA. The dendrogram (**Supplementary Figure S1**) showed

three main clusters (I to III), besides a few samples diverging from them (A and B). Main branches were supported by high-bootstrap values (>90). Moreover, bootstrap values were mainly high as well inside the main three clusters. Only some final subgroups had statistically non-significant bootstrap values. The separation in the dendrogram of some well-characterized samples (C1 to C5) is of special interest. Thus, Spanish varieties (Purple C3 and White C4; highlighted in purple and pink, respectively, in **Supplementary Figure S1**) were more related between them than to Chinese varieties (White C1 and Purple C2; highlighted in brown in **Supplementary Figure S1**), which were closely related. Sample C5 is a Brazilian garlic (thought to be an old Spanish Purple variety exported to America during colonialism) brought back to Spain 5 years ago. Interestingly, it was nearer to Spanish samples (closer to C3 than to C4) than to other accessions (C1 and C2), being highlighted in purple (**Supplementary Figure S1**).

Agro-morphological information (Supplementary Table S1) showed data in agreement with the generated dendrogram. For instance, cluster A contained samples 167, 239, and 459, being hexaploid or giant varieties (**Supplementary Figure S1**; highlighted with orange dots). There was a fourth hexaploid individual (379), being located in cluster III. Another interesting case was made of samples grouped together and with similar geographical origins. Thus, accessions 511, 513, and 514 came from Egypt (**Supplementary Figure S1**; highlighted with brown dots). Additionally, there were clusters with samples from Castilla-Leon region like: (i) 380, 389, and 432; (ii) 376, 424, 425, and 431; and (iii) 54, 423, 434, and 438 in the case of cluster II (highlighted with pink dots). Samples 32, 123, 125, 136, 225, and 1390 in cluster III were from Andalusia region (Spain; highlighted with purple dots). Samples 265, 270, 272 to 274, 276, 300, and 373 from cluster B came from Japan.

In addition, most accessions were also grouped by garlic-variety color in the phylogenetic tree. Thus, samples 20, 54, 238, 335, 360, 368, 424, 452, and 467 were Spanish White varieties (cluster II, pink). Likewise, samples 2, 3, 16, 17, 19, 21, 27, 29, 30, 32, 33, 37, 38, 77, 85, 87, 110, 117, 120, 123 to 125, 131, 132, 136, 138 to 141, 149, 150, 158, 161, 166, 171 to 173, 296, 297, 342, 343, 349, 356, 366, 454, 489, 542, 543, 560, 566, 570, 572, 574, 577, 578, 694, 752, 774, 779, G and K were Spanish Purple, Red, Brown, or "Colorado" varieties (cluster III, purple). Conversely, some samples did not group as expected. Thus, accessions 176 and 353 (Brown and Spanish Purple, respectively) would belong to cluster III, in accordance to their available agro-morphological data, yet they were in cluster A. Likewise, samples 36, 43, 88, and 109 (being considered Red or Purple varieties) did not group in cluster III, but in cluster II instead. Additionally, sample 44 is described as Chinese and thus expected in cluster I, but showed in cluster II instead. Samples 28, 79, 101, 137, 268, 526, 753, 776, and L (described as White varieties) were expected in cluster II, but were in cluster III. Sample 51 (described as Spanish White) was conversely located in cluster I instead of II. Likewise for some Spanish Purple samples (7, 348, 363, 369, and 775). Finally, samples 263 and 300 (described as White varieties) were included in cluster B instead of II. All samples that were not assigned consistently with agro-morphological data were highlighted with red dots in **Supplementary Figure S1**.

Principal-coordinates analysis was performed to further evaluate dendrogram clusters (**Figure 1**). Variance (genetic diversity) explained by principal components (PC) (accounting for 0.99 of cumulative variance) was 0.93 for PC1, 0.04 for PC2, and 0.02 for PC3. The relationships for samples C1 to C5 were similar to the ones in the dendrogram. As expected, samples C1 and C2 were nearer among them (Chinese), as well as samples C3 to C5 (Spanish origin). In addition, samples C3 and C5 were also closer compared to C4, as displayed in dendrogram (Supplementary Table S4).

Germplasm Genetic-Structure

Genetic structure of the garlic germplasm-bank collection was evaluated with STRUCTURE software. Three groups were

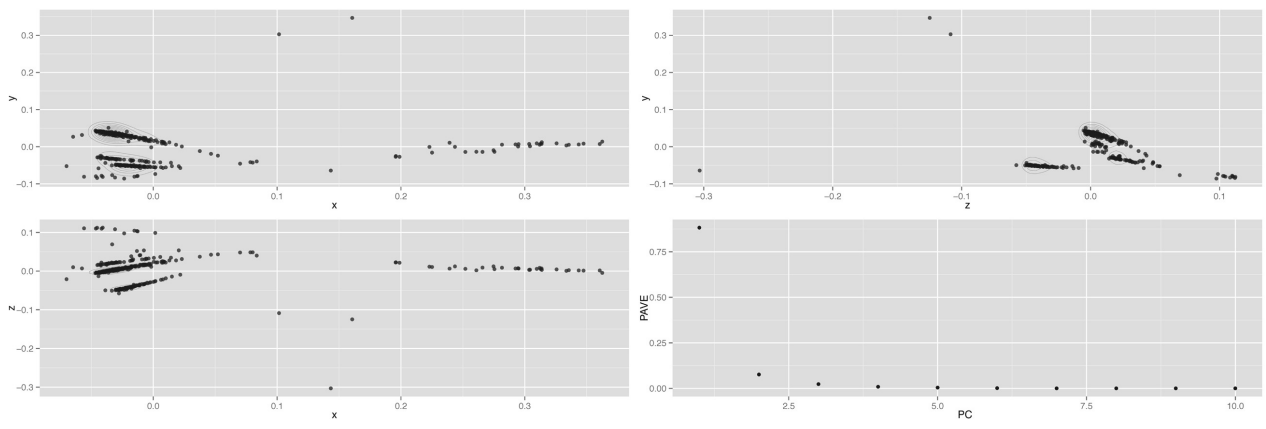


FIGURE 1 | Garlic principal-component plot. PCoA analysis was carried out to further analyze the garlic germplasm diversity. Upper-left represents principal coordinate (PC1; *x*-axis) with PC2 (*y*-axis); lower-left compares PC1 (*x*-axis) to PC3 (*z*-axis); and upper-right shows PC2 (*z*-axis) versus PC3 (*y*-axis). The lower-right graph corresponds to the Proportion of Analysis of Variance Explained (PAVE).

found, based on maximum likelihood and delta K (ΔK) values (**Supplementary Figure S2a**). As described above, this result is in agreement with cluster analysis and PCoA. Bar plot for $K = 3$ was also shown (**Supplementary Figure S2b**). In relation to the probability of membership of samples to clusters, Cluster I showed a score of 44.8%, being the group with the highest percentage. Clusters 2 and 3 had similar values (26.4 and 28.8%, respectively). When the probability of belonging to a group was high ($\leq$0.8 to 0.9), such individuals showed the same association found in hierarchical cluster-analysis. Well-known varieties (C1 to C5), also maintained the same relationships (Supplementary Table S5).

DISCUSSION

Garlic is known for multiple alimentary, medical and cosmetic uses worldwide. Yet, its classification and conservation in germplasm banks is challenging, due to homonymy and synonymy, being further complicated by its asexual life-cycle (Ipek et al., 2005). Previous information available allowed classifying the studied germplasm samples in this work by agro-morphological traits. Yet, such approach may be non-effective identifying true biodiversity, increasing redundancies and thus space and preservation costs in germplasm banks. In fact, it is known that the same garlic genotypes in different environmental conditions could exhibit diverse phenotypes (Volk et al., 2004). This is due to the high phenotypic plasticity of garlic, probably linked to its huge and expected complex genome, which somehow should compensate its lack of sexual reproduction.

Molecular markers have become an essential tool to identify, manage, and protect genetic diversity. Yet, developing them may be complicated, time-consuming and expensive for species like garlic, without sequenced reference genome, in which only scarce genomic-information is available (Ovesná et al., 2014). Additionally, classical molecular markers like isozymes, RAPD, SSR, or AFLP are not well suited to genotype garlic germplasm banks, due to its lack of resolution for such a peculiar genome in asexually reproducing accessions. Fortunately, technologies like DArT –and more recently, DArTseq– allow to reduce complexity and thus resolve complex genomic samples (Jaccoud et al., 2001).

Therefore, DArTseq was used in the present work to evaluate the genetic diversity and structure of 417 garlic samples (408 accessions from a garlic-germplasm bank). Data were analyzed by hierarchical-cluster computed by complete-linkage method, PCoA and genetic-structure approaches. Results showed a general consistency between accessions, geographic origins and groupings for expected/known garlic identities. All tests showed that individuals could be divided into three main groups (I, II, and III). Moreover, when the statistical probability of belonging to a group was high, the same association pattern of individuals was found in hierarchical-cluster analysis. Specifically, patterns for samples C1 to C5 (according to the previously known information) were maintained. Hence, DArTseq markers proved to be an effective and consistent genotyping approach to assess genetic diversity and structure.

Samples grouped by variety or geographical proximity were also found in non-redundant accessions, as described in the "Results" section. As expected, garlic samples of the same or near geographical regions grouped together. Indeed, cultivated varieties are usually selected by growers for several reasons, including being adapted to the climate in a specific region. In addition, the asexual garlic reproduction could lead to less genetic diversity and differentiation among varieties with similar geographical origins or different variants of the same variety. On the other hand, some samples were not grouped as expected, according to their agro-morphological information. Yet, such data is generated *de visu*, being therefore less accurate than molecular studies. In fact, it is known that morphological data are not always reliable to classify and detect genetic variation in germplasm collections (Jansky et al., 2015).

On the other hand, STRUCTURE assumes that markers are not in linkage disequilibrium (LD) within subpopulations. Yet, there are redundant lines in the data set, which could be against such assumption. But, there was a high consistency when comparing dendrogram clusters with those generated by STRUCTURE software. Thus, individuals assigned to the same cluster in the former, usually had higher probabilities to belong to the same group in the latter. Only three individuals were assigned differently in such analyses (4, 43, and 430) (Supplementary Table S5 and **Supplementary Figure S2**). This could be due to several reasons. In fact, criteria and calculations could lead to different results in each analysis. In the case of samples 4 and 430, they were located in an initial branch of cluster III, which indicates that they were genetically more different that the rest of assigned samples. Additionally, agro-morphological information was missing for samples 4 and 430.

The redundancy analysis showed that about one third of studied samples (131) could be considered as genetically redundant vs. 286 non-redundant (unique). This shows the higher resolution power and value of genomic analyses over agro-morphological ones. Thus, DArTseq results allowed to significantly reduce the analyzed garlic germplasm-bank size by 31.41%, generating a core collection, which was the main purpose of this research. Redundant accessions were divided into 19 groups (Supplementary Table S3). Samples included in each of them were in general related by variety (White, Purple, etc.) or location (same or near provinces). Interestingly, White varieties were more differentiated by location, whereas Purple ones were mainly associated in only one group. Samples 79 (Chinese White variety) and 526 (Spanish White variety) showed in group 7, in which Spanish Purple individuals were included. Curiously, this same lack of correlation was found in the hierarchical-cluster analysis, suggesting identities/differences not yet well understood. Further research is required to properly assess such results, including analyses of full genome sequences, once available in the future. That is now a possibility for large genomes like the garlic one, thanks to the throughput increase and cost reduction of TGS, which is expected to

become a mature technology in the next years (Dorado et al., 2015b).

As we have found, DArTseq is a cost-effective genotyping tool for creating and maintaining germplasm banks, allowing to properly ascertain, manage and maintain available biodiversity. Such technology has generated high-quality whole-genome profiles and genetic patterns, with dramatically increased resolution in relation to previous methodologies. Additionally, the high number of samples analyzed in this work, together with the large amount of marker data generated on lines with phenotypic information, should be useful for both genetic dissection of important traits and to help breeders improve this crop. Moreover, results obtained by DArTseq in any species can help to perform further analyses in germplasm collections without previous genetic information, even with high phenotypic-plasticity, complex genomes and asexual reproductive-systems that may hamper diversity analyses (Gebhardt, 2013). DArTseq sequences can be used to develop DArTseq markers and other molecular markers, such as SSR or SNP, which can be transferable to other germplasm banks (Belaj et al., 2011; Atienza et al., 2013). These tools can be associated to traits of interest, and thus used for marker-assisted breeding.

CONCLUSION

We have significantly reduced the analyzed garlic germplasm-bank size, identifying redundant accessions and thus generating a unique (non-redundant) core collection, with the consequent reduction in space and maintenance expenses. To our knowledge, this is the first work of high-throughput garlic genotyping. The obtained results show that DArTseq is a cost-effective method to perform genotyping-by-sequencing and genetic diversity analyses of such species with huge, expected complex and mostly unknown (without reference) genome, with clear applications for biodiversity conservation. This supports previous studies for characterizing and managing germplasm banks of other species. DArTseq has generated consistent results, in accordance with variety and geographical origin. They remark the relevance of genetic versus agro-morphological data, especially in the context of peculiar garlic-plasticity for environmental adaptation. Additionally, the high number of samples analyzed in this work and the amount of data generated should be useful for plant breeders in general, as well as for garlic adaptation and improvement in particular. This, along with other molecular markers and agro-morphological information represent useful tools to improve management strategies in germplasm-banks. In fact, having a core collection of characterized genotypes and phenotypes could help breeders to select plants with better adaptability. This is important for productivity and to face biotic and abiotic stresses, to fight the current climate change and global warming.

AUTHOR CONTRIBUTIONS

LE performed experiments, analyzed data, and wrote the manuscript; RM-G analyzed data and participated in manuscript writing; AK contributed to reagents, analysis tools, and manuscript writing; PH contributed to experimental design, materials, reagents, analysis tools, and manuscript writing; GD conceived and designed the experiments, contributed to materials, reagents, analysis tools, and manuscript writing; All authors read and approved the final version of the manuscript.

FUNDING

Supported by "Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria" (MINECO and INIA RF2012-00002-C02-02) and jointly funded by "Fondo Europeo de Desarrollo Regional" (FEDER); "Consejería de Agricultura y Pesca" (041/C/2007, 75/C/2009 and 56/C/2010), "Consejería de Economía, Innovación y Ciencia" (P11-AGR-7322) and "Grupo PAI" (AGR-248) of "Junta de Andalucía"; and "Universidad de Córdoba" ("Ayuda a Grupos"), Spain.

ACKNOWLEDGMENTS

We thank Francisco Mansilla (IFAPA, Córdoba, Spain) for germplasm samples and agro-morphological data. Likewise, Jesús Martín and Jaime Martín ("Universidad de Córdoba"; and "Innovolivo", Córdoba, Spain) and Antonio Escolano ("Centro de Ensayos de Evaluación de Variedades", INIA, Madrid) for additional garlic-samples. Teresa Hernández-Gutiérrez is acknowledged for support during sampling and other experimental work, and Jaroslava Ovesná (Crop Research Institute, Prague, Czechia) for comments on garlic genotyping.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00098/full#supplementary-material

FIGURE S1 | Garlic dendrogram. Phylogenetic tree, with approximately unbiased (AU; red)/Bootstrap Probability (BP; green) percentage values and Euclidean distances, generated by complete-linkage method, to ascertain germplasm diversity. Cluster I includes C1 and C2 (Chinese varieties); Cluster II has C4 (Spanish White variety); and Cluster III shows C3 to C5 (Spanish Purple and Brazilian varieties). Samples C1 to C5, and others described in the text, are highlighted with colored dots. I corresponds to cluster II in STRUCTURE analysis, whereas II and III are equivalent to cluster I; and A and B correspond to cluster III using such software analysis.

FIGURE S2 | Garlic genetic structure. STRUCTURE software was used to analyze the studied garlic germplasm. **(a)** Diagram showing the three calculated clusters ($K = 3$); and **(b)** ΔK values.

REFERENCES

Akpinar, B., Lucas, S., and Budak, H. (2017). A large-scale chromosome-specific SNP discovery guideline. *Funct. Integr. Genom.* 17, 97–105. doi: 10.1007/s10142-016-0536-6

Al-Zahim, M., Ford-Lloyd, B., and Newbury, H. (1999). Detection of somaclonal variation in garlic (*Allium sativum* L.) using RAPD and cytological analysis. *Plant Cell Rep.* 18, 473–477. doi: 10.1007/s002990050606

Arumuganathan, K., and Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* 9, 208–218. doi: 10.1007/BF02672069

Atienza, S. G., de la Rosa, R., Domínguez-García, M. C., Martín, A., Kilian, A., and Belaj, A. (2013). Use of DArT markers as a means of better management of the diversity of olive cultivars. *Food Res. Int.* 54, 2045–2053. doi: 10.1016/j.foodres.2013.08.015

Belaj, A., Dominguez-García, M. D. C., Atienza, S. G., Urdíroz, N. M., Rosa, R. D., Satovic, Z., et al. (2011). Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genet. Genomes* 8, 365–378. doi: 10.1007/s11295-011-0447-6

Cardelle-Cobas, A., Soria, A. C., Corzo-Martinez, M., and Villamiel, M. (2010). "A comprehensive survey of garlic functionality," in *Garlic Consumption and Health*, eds M. Pacurar and G. Krejci (Hauppauge: Nova Science Publishers, Inc), 1–60.

Carlsson, K., and Frick, G. (1964). Partial purification of nucleases from germinating garlic. *Biochim. Biophys. Acta* 81, 301–310. doi: 10.1016/0926-6569(64)90046-x

Corzo-Martínez, M., Corzo, N., and Villamiel, M. (2007). Biological properties of onions and garlic. *Trends Food Sci. Technol.* 18, 609–625. doi: 10.1016/j.tifs.2007.07.011

Courtois, B., Audebert, A., Dardou, A., Roques, S., Ghneim-Herrera, T., Droc, G., et al. (2013). Genome-wide association mapping of root traits in a japonica rice panel. *PLoS ONE* 8:e78037. doi: 10.1371/journal.pone.0078037

Cruz, V. M. V., Kilian, A., and Dierig, D. A. (2013). Development of DArT marker platforms and genetic diversity assessment of the US collection of the new oilseed crop lesquerella and related species. *PLoS ONE* 8:e64062. doi: 10.1371/journal.pone.0064062

DaCunha, C. P., Resende, F. V., Zucchi, M. I., and Pinheiro, J. B. (2014). SSR-based genetic diversity and structure of garlic accessions from Brazil. *Genetica* 142, 419–431. doi: 10.1007/s10709-014-9786-1

Defays, D. (1977). Efficient algorithm for a complete link method. *Comput. J.* 20, 364–366. doi: 10.1093/comjnl/20.4.364

Dorado, G., Besnard, G., Unver, T., and Hernández, P. (2015a). "Polymerase chain reaction (PCR)," in *Reference Module in Biomedical Sciences*, ed. M. Caplan (Amsterdam: Elsevier).

Dorado, G., Gálvez, S., Budak, H., Unver, T., and Hernández, P. (2015b). "Nucleic-acid sequencing," in *Reference Module in Biomedical Sciences*, ed. M. Caplan (Amsterdam: Elsevier).

Dorado, G., Unver, T., Budak, H., and Hernández, P. (2015c). "Molecular markers," in *Reference Module in Biomedical Sciences*, ed. M. Caplan (Amsterdam: Elsevier).

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0054603

Gebhardt, C. (2013). Bridging the gap between genome analysis and precision breeding in potato. *Trends Genet.* 29, 248–256. doi: 10.1016/j.tig.2012.11.006

Govindaraj, M., Vetriventhan, M., and Srinivasan, M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genet. Res. Int.* 2015, 431487–431487. doi: 10.1155/2015/431487

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27, 857–871. doi: 10.2307/2528823

Green, E. (2001). Strategies for the systematic sequencing of complex genomes. *Nat. Rev. Genet.* 2, 573–583. doi: 10.1038/35084503

Gupta, P. K., Rustgi, S., and Mir, R. R. (2008). Array-based high-throughput DNA markers for crop improvement. *Heredity* 101, 5–18. doi: 10.1038/hdy.2008.35

He, N., Li, Q., Sun, D., and Ling, X. (2008). Isolation, purification and characterization of superoxide dismutase from garlic. *Biochem. Eng. J.* 38, 33–38. doi: 10.1016/j.bej.2007.06.005

Hernandez, P., de la Rosa, R., Rallo, L., Martin, A., and Dorado, G. (2001). First evidence of a retrotransposon-like element in olive (*Olea europaea*): implications in plant variety identification by SCAR-marker development. *Theor. Appl. Genet.* 102, 1082–1087. doi: 10.1007/s001220000515

Ipek, A., Yilmaz, K., Sikici, P., Tangu, N., Oz, A., Bayraktar, M., et al. (2016). SNP discovery by GBS in olive and the construction of a high-density genetic linkage map. *Biochem. Genet.* 54, 313–325. doi: 10.1007/s10528-016-9721-5

Ipek, M., Ipek, A., Almquist, S. G., and Simon, P. W. (2005). Demonstration of linkage and development of the first low-density genetic map of garlic, based on AFLP markers. *Theor. Appl. Genet.* 110, 228–236. doi: 10.1007/s00122-004-1815-5

Ipek, M., Sahin, N., Ipek, A., Cansev, A., and Simon, P. (2015). Development and validation of new SSR markers from expressed regions in the garlic genome. *Sci. Agric.* 72, 41–46. doi: 10.1590/0103-9016-2014-0138

Jaccoud, D., Peng, K., Feinstein, D., and Kilian, A. (2001). Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res.* 29:E25. doi: 10.1093/nar/29.4.e25

Jansky, S. H., Dawson, J., and Spooner, D. M. (2015). How do we address the disconnect between genetic and morphological diversity in germplasm collections? *Am. J. Bot.* 102, 1213–1215. doi: 10.3732/ajb.1500203

Jones, M. G., Hughes, J., Tregova, A., Milne, J., Tomsett, A. B., and Collin, H. A. (2004). Biosynthesis of the flavour precursors of onion and garlic. *J. Exp. Bot.* 55, 1903–1918. doi: 10.1093/jxb/erh138

Kamenetsky, R., Faigenboim, A., Mayer, E., Ben Michael, T., Gershberg, C., Kimhi, S., et al. (2015). Integrated transcriptome catalogue and organ-specific profiling of gene expression in fertile garlic (*Allium sativum* L.). *BMC Genomics* 16:12. doi: 10.1186/s12864-015-1212-2

Kamenetsky, R., Shafir, I. L., Baizerman, M., Khassanov, F., Kik, C., and Rabinowitch, H. D. (2004). Garlic (*Allium sativum* L.) and its wild relatives from Central Asia: evaluation for fertility potential. *Adv. Vegetable Breed.* 83–91.

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., et al. (2012). Diversity arrays technology: a generic genome profiling technology on open platforms. *Methods Mol. Biol. (Clifton, NJ)* 888, 67–89. doi: 10.1007/978-1-61779-870-2_5

Kim, A., Kim, R. N., Kim, D.-W., Choi, S-H., Kang, A., Nam, S.-H., et al. (2010). Identification of a novel garlic cellulase gene. *Plant Mol. Biol. Rep.* 28, 388–393. doi: 10.1007/s11105-009-0159-3

Kim, D.-W., Jung, T.-S., Nam, S.-H., Kwon, H.-R., Kim, A., Chae, S.-H., et al. (2009). GarlicESTdb: an online database and mining tool for garlic EST sequences. *BMC Plant Biol.* 9:61. doi: 10.1186/1471-2229-9-61

Liu, J., Wang, J., Yin, M., Zhu, H., Lu, J., and Cui, Z. (2011). Purification and characterization of superoxide dismutase from garlic. *Food Bioprod. Process.* 89, 294–299. doi: 10.1016/j.fbp.2010.07.003

Maab, H. I., and Klaas, M. (1995). Infraspecific differentiation of garlic (*Allium sativum* L.) by isozyme and RAPD markers. *Theor. Appl. Genet.* 91, 89–97.

Meredith, T. (2008). *The Complete Book of Garlic: A Guide for Gardeners, Growers, and Serious Cooks*. Portland: Timber Press.

Murray, M. G., and Thompson, W. F. (1980). Rapid isolation of high molecular-weight plant DNA. *Nucleic Acids Res.* 8, 4321–4325. doi: 10.1093/nar/8.19.4321

Nybom, H., and Bartish, I. (2000). Effects of life history traits and sampling strategies on genetic diversity estimates obtained with RAPD markers in plants. *Perspect. Plant Ecol. Evol. Syst.* 3, 93–114. doi: 10.1078/1433-8319-00006

Olukolu, B. A., Mayes, S., Stadler, F., Ng, N. Q., Fawole, I., Dominique, D., et al. (2012). Genetic diversity in *Bambara groundnut* (*Vigna subterranea* (L.) Verdc.) as revealed by phenotypic descriptors and DArT marker analysis. *Genet. Res. Crop Evol.* 59, 347–358. doi: 10.1007/s10722-011-9686-5

Ovesná, J., Leišová-Svobodová, L., and Kučera, L. (2014). Microsatellite analysis indicates the specific genetic basis of Czech bolting garlic. *Czech J. Genet. Plant Breed.* 50, 226–234.

Ovesna, J., Mitrova, K., and Kucera, L. (2015). Garlic (*A. sativum* L.) alliinase gene family polymorphism reflects bolting types and cysteine sulphoxides content. *BMC Genet.* 16:53. doi: 10.1186/s12863-015-0214-z

Pacurar, M., and Krejci, G. (eds). (2010). *Garlic Consumption and Health*. New York, NY: Nova Science Publishers.

Parisi, M., Moreno, S., and Fernandez, G. (2008). Isolation and characterization of a dual function protein from *Allium sativum* bulbs which exhibits proteolytic and hemagglutinating activities. *Plant Physiol. Biochem.* 46, 403–413. doi: 10.1016/j.plaphy.2007.11.003

Park, J. (2009). Isolation and characterization of N-Feruloyltyramine as the P-selectin expression suppressor from garlic (*Allium sativum*). *J. Agric. Food Chem.* 57, 8868–8872. doi: 10.1021/jf9018382

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

R-Development-Core-Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raman, H., Raman, R., Kilian, A., Detering, F., Carling, J., Coombes, N., et al. (2014). Genome-wide delineation of natural variation for pod shatter resistance in *Brassica napus*. *PLoS ONE* 9:e101673. doi: 10.1371/journal.pone.0101673

Rana, S., Pal, R., Vaiphei, K., Sharma, S., and Ola, R. (2011). Garlic in health and disease. *Nutr. Res. Rev.* 24, 60–71. doi: 10.1017/S0954422410000338

Rathnasamy, S., Auxilia, L. R., and Purusothaman. (2014). Comparative studies on isolation and characterization of allinase from garlic and onion using PEGylation-a novel method. *Asian J. Chem.* 26, 3733–3735.

Rotem, N., Shemesh, E., Peretz, Y., Akad, F., Edelbaum, O., Rabinowitch, H., et al. (2007). Reproductive development and phenotypic differences in garlic are associated with expression and splicing of LEAFY homologue gaLFY. *J. Exp. Bot.* 58, 1133–1141. doi: 10.1093/jxb/erl272

Shemesh-Mayer, E., Ben-Michae, T., Rotem, N., Rabinowitch, H., Doron-Faigenboim, A., Kosmala, A., et al. (2015*). Garlic (*Allium sativum* L.) fertility: transcriptome and proteome analyses provide insight into flower and pollen development. *Front. Plant Sci.* 6:271. doi: 10.3389/fpls.2015.00271

Smeets, K., Van Damme, E., Van Leuven, F., and Peumans, W. (1997). Isolation and characterization of lectins and lectin-alliinase complexes from bulbs of garlic (*Allium sativum*) and ramsons (*Allium ursinum*). *Glycoconj. J.* 14, 331–343. doi: 10.1023/A:1018570628180

Suetsuna, K. (1998). Isolation and characterization of angiotensin I-converting enzyme inhibitor dipeptides derived from *Allium sativum* L (garlic). *J. Nutr. Biochem.* 9, 415–419. doi: 10.1016/S0955-2863(98)00036-9

Suzuki, R., and Shimodaira, H. (2015). *pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling*. Available at: http://stat.sys.i.kyoto-u.ac.jp/prog/pvclust/

Tanksley, S. D., and McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277, 1063–1066. doi: 10.1126/science.277.5329.1063

Volk, G. M., Henk, A. D., and Richards, C. M. (2004). Genetic diversity among U.S. Garlic clones as detected using AFLP methods. *J. Am. Soc. Hortic. Sci.* 129, 559–569.

Wang, H., Li, X., Liu, X., Oiu, Y., Song, J., and Zhang, X. (2016). Genetic diversity of garlic (*Allium sativum* L.) germplasm from China by fluorescent-based AFLP, SSR and InDel markers. *Plant Breed.* 135, 743–750. doi: 10.1111/pbr.12424

Zhao, W. G., Chung, J. W., Lee, G. A., Ma, K. H., Kim, H. H., Kim, K. T., et al. (2010). Molecular genetic diversity and population structure of a selected core set in garlic and its relatives using novel SSR markers. *Plant Breed.* 130, 46–54. doi: 10.1111/j.1439-0523.2010.01805.x

Conflict of Interest Statement: AK works at Diversity Arrays Technology. This fact did not interfere with the objective, transparent and unbiased presentation of results, and does not alter the authors' adherence to all theoretical and applied genetics policies on data and material release.

The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

1.8. Supplementary Material

Supplementary Material 1.1. Analyzed garlic accessions (417) *.

ID	Taxon	Description	Origin
1	*Allium sativum*	–	Cabeza del Obispo (Sevilla, Spain)
2	*Allium sativum*	Spanish Purple	Pedroñeras (Cuenca, Spain)
3	*Allium sativum*	Spanish Purple	Cordoba (Spain)
7	*Allium sativum*	Purple	France
14	*Allium sativum*	–	Taiwan
16	*Allium sativum*	Spanish Purple	Setiles (Guadalajara, Spain)
17	*Allium sativum*	Spanish Purple	Simeu (Mallorca, Spain)
19	*Allium sativum*	Spanish Purple	Mahon (Mallorca, Spain)
20	*Allium sativum*	Spanish White	–
21	*Allium sativum*	Spanish Purple	Terrinches (Ciudad Real, Spain)
27	*Allium sativum*	Spanish Purple	Villarrubia (Cordoba, Spain)
28	*Allium sativum*	Spanish White	Villarrubia (Cordoba, Spain)
29	*Allium sativum*	Spanish Purple	Mondejar (Guadalajara, Spain)
30	*Allium sativum*	Pink	Almoguera (Guadalajara, Spain)
32	*Allium sativum*	Purple	Lanteira (Granada, Spain)
33	*Allium sativum*	Purple	Lanteira (Granada, Spain)
36	*Allium sativum*	Purple	–
37	*Allium sativum*	Spanish Purple	Terrinches (Ciudad Real, Spain)
38	*Allium sativum*	Spanish Purple	Alhambra (Ciudad Real, Spain)
39	*Allium sativum*	French White	–
41	*Allium sativum*	Ajofrin	–
43	*Allium sativum*	Red	Bañolas (Gerona, Spain)
44	*Allium sativum*	Chinese	–
45	*Allium sativum*	Basic	–
47	*Allium sativum*	–	Morocco
50	*Allium sativum*	Christ	–
51	*Allium sativum*	Spanish White	Ronda (Malaga, Spain)
54	*Allium sativum*	Spanish White	Vallelado (Segovia, Spain)
59	*Allium sativum*	Spanish Purple	Pedroñeras (Cuenca, Spain)
74	*Allium* spp. (cultivated hexaploid)	Giant	Netherlands
76	*Allium sativum*	–	Antequera (Malaga, Spain)
77	*Allium sativum*	Red	Bañolas (Gerona, Spain)
78	*Allium sativum*	Basic	–
79	*Allium sativum*	Chinese White	–
85	*Allium sativum*	Red	Rute (Cordoba, Spain)
86	*Allium sativum*	Red	Bañolas (Gerona, Spain)
87	*Allium sativum*	Red	Castro (Cordoba, Spain)

88	*Allium sativum*	Red	Bañolas (Gerona, Spain)
89	*Allium sativum*	–	Pedroche (Cordoba, Spain)
91	*Allium sativum*	Seversky Palicak	–
92	*Allium sativum*	Creole	–
96	*Allium sativum*	Ophioscorodon	–
98	*Allium sativum*	Adizanskij	–
100	*Allium sativum*	Chines	–
101	*Allium sativum*	Spanish White	Huelma (Jaen, Spain)
102	*Allium sativum*	Spanish White	Huelma (Jaen, Spain)
107	*Allium sativum*	Spanish Purple	Jaen (Spain)
109	*Allium sativum*	Red	Falces (Navarre, Spain)
110	*Allium sativum*	Red	La Carlota (Cordoba, Spain)
114	*Allium sativum*	–	Villatoya (Albacete, Spain)
116	*Allium sativum*	–	Galicia (Spain)
117	*Allium sativum*	Brown	Salamanca (Spain)
119	*Allium sativum*	–	La Serena (Badajoz, Spain)
120	*Allium sativum*	Spanish Purple	Olula del Rio (Almeria, Spain)
123	*Allium sativum*	Spanish Purple	Padules (Almeria, Spain)
124	*Allium sativum*	Spanish Purple	Tabernas (Almeria, Spain)
125	*Allium sativum*	Spanish Purple	Padul (Granada, Spain)
126	*Allium sativum*	Kamara	–
127	*Allium sativum*	Spanish White	Torrecampo (Cordoba, Spain)
130	*Allium sativum*	Salvador	–
131	*Allium sativum*	Spanish Purple	Gador (Almeria, Spain)
132	*Allium sativum*	Spanish Purple	Fuente Victoria (Almeria, Spain)
136	*Allium sativum*	Spanish Purple	Nigüelas (Granada, Spain)
137	*Allium sativum*	Spanish White	Hinojosa del Duque (Cordoba, Spain)
138	*Allium sativum*	Spanish Purple	Monturque (Cordoba)
139	*Allium sativum*	Spanish Purple	Acequias (Granada, Spain)
140	*Allium sativum*	Spanish Purple	Olula del Rio (Almeria, Spain)
141	*Allium sativum*	Spanish Purple	Purullena (Granada, Spain)
149	*Allium sativum*	Spanish Purple	Padul (Granada, Spain)
150	*Allium sativum*	Spanish Purple	Acequias (Granada, Spain)
151	*Allium sativum*	Spanish Purple	Santa Cruz (Cordoba, Spain)
152	*Allium sativum*	Spanish Purple	Trevelez (Granada, Spain)
158	*Allium sativum*	Spanish Purple	Ugijar (Granada, Spain)
161	*Allium sativum*	Spanish Purple	–
162	*Allium sativum*	Cardenal Liv	–
166	*Allium sativum*	Spanish Purple	–
167	*Allium* spp. (cultivated hexaploid)	Giant	Navas de la Concepcion
171	*Allium sativum*	Brown	Castilleja de la Cuesta (Sevilla, Spain)
172	*Allium sativum*	Brown	Corteconcepcion (Huelva, Spain)
173	*Allium sativum*	Brown	Santa Ana la Real (Huelva, Spain)
176	*Allium sativum*	Brown	Cerro de Andevalo (Huelva, Spain)

189	*Allium* spp.	–	Castilleja de la Cuesta (Sevilla, Spain)
193	*Allium* spp.	–	La Carlota (Cordoba, Spain)
217	*Allium* spp.	–	Villanueva de la Reina (Jaen, Spain)
219	*Allium* spp.	–	La Carlota (Cordoba, Spain)
225	*Allium* spp.	–	Estepona (Malaga)
238	*Allium sativum*	Spanish White	Mondoñedo (Lugo, Spain)
239	*Allium* spp. (cultivated hexaploid)	Giant	Ribadeo (Lugo, Spain)
243	*Allium sativum*	–	Santa Comba (La Coruña, Spain)
246	*Allium sativum*	Pink	Tegucigalpa (Honduras)
263	*Allium sativum*	White Roppen	–
265	*Allium sativum*	Nigata Sado	–
266	*Allium sativum*	Ibaraki	–
268	*Allium sativum*	White Roppen	–
270	*Allium sativum*	Hamamtsu	–
272	*Allium sativum*	Shimane Tsunozu	–
273	*Allium sativum*	Kochi Daikyu	–
274	*Allium sativum*	Kagoshima	–
276	*Allium sativum*	Toroku-Kuroba-Kokotsu	–
278	*Allium sativum*	Bansei	–
280	*Allium sativum*	–	Taiwan
296	*Allium sativum*	Red	Peñalsordo (Badajoz, Spain)
297	*Allium sativum*	Spanish Purple	El Vacar (Cordoba, Spain)
299	*Allium sativum*	Talca	–
300	*Allium sativum*	Spanish White	Mallorca (Spain)
328	*Allium sativum*	Arica	–
332	*Allium sativum*	Mabegondo	–
334	*Allium sativum*	Vallemar	–
335	*Allium sativum*	Spanish White	Puenteviesgo (Cantabria, Spain)
338	*Allium sativum*	Thermidrome	–
339	*Allium sativum*	Fructidor	–
342	*Allium sativum*	Spanish Purple	Cazorla (Jaen, Spain)
343	*Allium sativum*	Spanish Purple	Laujar de Andarax (Almeria, Spain)
344	*Allium sativum*	Spanish Purple	Albox (Almeria, Spain)
348	*Allium sativum*	Spanish Purple	Velez Rubio (Almeria, Spain)
349	*Allium sativum*	Spanish Purple	Velez Rubio (Almeria, Spain)
353	*Allium sativum*	Spanish Purple	Mula (Murcia, Spain)
354	*Allium* spp.	Giant	Alhama (Granda, Spain)
356	*Allium sativum*	Spanish Purple	Barranda (Murcia, Spain)
360	*Allium sativum*	Spanish White	Campanario (Badajoz, Spain)
363	*Allium sativum*	Spanish Purple	Guadalupe (Caceres, Spain)
364	*Allium* spp.	Allium porrum	Guadalupe (Caceres, Spain)
366	*Allium sativum*	Spanish Purple	Valdecaballeros (Badajoz, Spain)
367	*Allium sativum*	Chinese Purple	Valdecaballeros (Badajoz, Spain)
368	*Allium sativum*	Spanish White	Navalmoral de la Mata (Caceres, Spain)

369	*Allium sativum*	Spanish Purple	Navalmoral de la Mata (Caceres, Spain)
373	*Allium sativum*	Sendai	–
376	*Allium sativum*	–	Alaejos (Valladolid, Spain)
377	*Allium sativum*	Puentenuevo	–
379	*Allium* spp. (cultivated hexaploid)	Chilote	–
380	*Allium sativum*	–	Cubo de Don Sancho (Salamanca, Spain)
384	*Allium sativum*	–	Argujillo (Zamora, Spain)
386	*Allium sativum*	–	Casas del Conde (Salamanca, Spain)
389	*Allium sativum*	–	Tolilla (Zamora, Spain)
390	*Allium sativum*	–	Gergal (Almeria, Spain)
403	*Allium sativum*	Rusuli Niri	–
404	*Allium sativum*	Bogatyr	–
409	*Allium sativum*	Kartuli Niori	–
418	*Allium sativum*	–	Pedraja de Portillo (Valladolid, Spain)
423	*Allium sativum*	–	La Santa Espina (Valladolid, Spain)
424	*Allium sativum*	Spanish White	Alaejos (Valladolid, Spain)
425	*Allium sativum*	–	Fuentelapeña (Zamora, Spain)
431	*Allium sativum*	–	Zamora (Spain)
432	*Allium sativum*	–	Zamora (Spain)
433	*Allium sativum*	–	Zamora (Spain)
434	*Allium sativum*	–	Zamora (Spain)
438	*Allium sativum*	–	Zamora (Spain)
440	*Allium sativum*	–	Zamora (Spain)
444	*Allium sativum*	–	Leon (Spain)
452	*Allium sativum*	Spanish White	Monzon (Huesca, Spain)
454	*Allium sativum*	Red	Yegen (Granada, Spain)
457	*Allium sativum*	–	Rañeces (Asturias, Spain)
459	*Allium sativum*	Giant	–
461	*Allium sativum*	–	Tineo (Asturias, Spain)
464	*Allium sativum*	–	Pola de Siero (Asturias, Spain)
467	*Allium sativum*	Spanish White	Libardon (Asturias, Spain)
469	*Allium sativum*	–	Betanzos (Asturias, Spain)
470	*Allium sativum*	–	Betanzos (Asturias, Spain)
486	*Allium sativum*	–	Teruel (Spain)
487	*Allium sativum*	–	Teruel (Spain)
489	*Allium sativum*	–	Teruel (Spain)
491	*Allium sativum*	–	Teruel (Spain)
494	*Allium sativum*	–	Teruel (Spain)
497	*Allium sativum*	–	Teruel (Spain)
502	*Allium sativum*	–	Zaragoza (Spain)
504	*Allium sativum*	–	Zaragoza (Spain)
506	*Allium sativum*	–	Egypt
510	*Allium sativum*	–	Egypt
511	*Allium sativum*	–	Egypt

513	*Allium sativum*	–	Egypt
514	*Allium sativum*	–	Egypt
517	*Allium sativum*	–	Italy
520	*Allium sativum*	–	Susanville (California, USA)
522	*Allium sativum*	Organic Elephant	–
523	*Allium sativum*	Soft Neck	Poland
526	*Allium sativum*	Premium White	–
533	*Allium sativum*	Spanish White	Mendoza (Alava, Spain)
536	*Allium sativum*	Messidrome	–
540	*Allium sativum*	–	Hungary
541	*Allium sativum*	Don Rafael	–
542	*Allium sativum*	Red	–
543	*Allium sativum*	Red	–
545	*Allium sativum*	Red	–
547	*Allium sativum*	–	Russia
550	*Allium sativum*	Cazador	–
553	*Allium sativum*	–	Santa Cruz (Cordoba, Spain)
556	*Allium sativum*	Nevado	–
559	*Allium sativum*	Inco	–
566	*Allium sativum*	Red	Falkland Islands (UK)
568	*Allium sativum*	Colorado	–
570	*Allium sativum*	Colorado	–
572	*Allium sativum*	Colorado	–
574	*Allium sativum*	Colorado	–
577	*Allium sativum*	Colorado	–
578	*Allium sativum*	Colorado	–
582	*Allium sativum*	Southern	–
583	*Allium sativum*	Northern	–
584	*Allium sativum*	Gostoso	–
585	*Allium sativum*	Chonan	–
592	*Allium sativum*	–	Mondoñedo (Lugo, Spain)
694	*Allium sativum*	Spanish Purple	Pedroñeras (Cuenca, Spain)
715	*Allium sativum*	Spanish Purple	Bayarcal (Almeria, Spain)
716	*Allium* spp.	Giant	Frailes (Jaen, Spain)
717	*Allium sativum*	–	Cabra de San Cristo (Jaen, Spain)
718	*Allium sativum*	–	Albanchez de Magina (Jaen, Spain)
720	*Allium sativum*	–	Bedmar (Jaen, Spain)
722	*Allium sativum*	–	Belmez de la Moraleda (Jaen, Spain)
750	*Allium sativum*	Purple	–
752	*Allium sativum*	Purple	–
753	*Allium sativum*	White	–
774	*Allium sativum*	Red	–
775	*Allium sativum*	Red	–
776	*Allium sativum*	White	–

779	*Allium sativum*	Red	–
C1	*Allium sativum*	Chinese white	–
C2	*Allium sativum*	Chinese purple	–
C3	*Allium sativum*	Spanish purple	–
C4	*Allium sativum*	Spanish white	–
C5	*Allium sativum*	Brazilian	–
G	*Allium sativum*	Spanish purple	–
K	*Allium sativum*	Spanish purple	–
L	*Allium sativum*	Spanish white	–

* Accessions without information were not included in the list (4, 10, 12, 13, 26, 64, 71, 90, 92, 93, 301 to 303, 314, 315, 324, 391, 394, 396, 427, 429, 430, 449, 466, 530, 531, 537, 587, 588, 590, 591, 593, 595, 596, 598, 600, 603 to 605, 607, 609, 614 to 619, 627, 630, 633, 634 to 637, 641 to 643, 651, 652, 654, 658 to 661, 664, 669, 670, 672 to 676, 680, 684, 685, 687, 688, 693, 696 to 698, 701 to 703, 705, 707 to 709, 711, 713, 723 to 727, 729, 731 to 733, 735, 736, 739, 742, 744, 745, 747, 754, 757, 760, 762, 763, 767, 769, 770 to 772, 777, 778, 780, 781, 783, 785, 787 to 789, 793, 800, 802, 804 to 816, 821 to 824, 827, 831 to 835, 837-838, 840, 842-843, 845 to 847, 871 to 879, 893, 900 to 903, 905, 907, 908-909, 911, 950 to 953, 955-956, 958 to 960, 962 to 966,1000 and M).

Supplementary Material 1.2. Unique garlic accessions

ID	ID	ID	ID	ID	ID	ID
1	141	363	556	713	807	907
4	149	364	559*	716	808	908
7	154	367	566	722	809	909
10	161	369	570	724	810	911
12	167*	373	572	725	811	950
26	173	376	574	729	812	951
27	176	377*	577	731	813	952
28	189	386	578	732	814	953
29	193	389	582*	739	815	955
30	217	391	583	745	816	956
33	219	394	584	747	821	958
36	239	396	588	750	822	959
39	243	403	591	752	823	960
41	246	404	592	753	824	962
43	249	409	593*	754	827	963
44	263	418*	595	757	831	964
45	266	423	599	760	832	965
47	270	424*	600	762	833	966
50	272	425*	605*	763	834	1000
51	273	427*	607*	767	835	
54	274	430	618	769	837	
74	276*	432	630	770	838	
76	278	440	636	771	840	
78	296	444	643	772	842	
86	297	449	658	774	843	
87	299*	452*	659	775	845	
88	300	459	660	776	846	
89	301	469	661	777	847	
91	302	487	664	778	871	
92	303	491	669	779	872	
96	315	494	670	780	873	
98	324*	502	672	781	874	
100	328	504	673	783	875	
101	332	506	674	785	876	
109	334	510	680	787	877	
114	339	511	684	788	878	
119	342	514	685	789	879	
120*	343*	520	687	793	893	
123	348	523	688	800	900	
126	353	530	697	802	901	
136	354	542	698	804	902	
137	356*	543	702	805	903	
140	358	550*	703	806	905	

*Randomly sample chosen to be kept from each of the 19 redundant groups.

Supplementary Material 1.3. Redundant garlic accessions

ID	ID	ID	ID
GROUP 1	**GROUP 6**	**GROUP 7**	**GROUP 10**
735	641	21	324*
377*	418*	132	497
717	16	526	124
675	540	585	162
718	338	547	64
676	633	617	568
723	314	489	694
GROUP 2	651	619	**GROUP 11**
90	**GROUP 7**	696	593*
434	615	708	596
127	150	720	**GROUP 12**
424*	225	711	550*
467	344	733	627
335	356*	736	742
20	280	85	701
553	541	107	**GROUP 13**
368	598	131	605*
457	642	366	609
536	110	705	**GROUP 14**
71	17	709	120*
634	252	**GROUP 8**	587
433	13	172	**GROUP 15**
470	654	427*	268
130	486	429	607*
360	545	652	517
531	744	590	116
464	715	156	**GROUP 16**
727	171	158	582*
GROUP 3	79	454	537
238	139	379	693
452*	614	604	**GROUP 17**
466	637	635	299*
461	603	19	513
707	14	349	726
GROUP 4	2	390	**GROUP 18**
431	59	**GROUP 9**	276*
425*	3	138	265
380	125	32	**GROUP 19**
438	37	343*	167*
GROUP 5	166	616	522
559*	38		
533	77	117	

*Randomly sample chosen to be kept from each of the 19 redundant groups.

Supplementary Material 1.4. PCoA coordinates.

ID	PC1	PC2	PC3
1	-0.03781317	0.03932796	0.000355133
2	-0.04054558	0.0395905	-0.002475114
3	-0.04400734	0.04139642	-0.004027278
4	-0.06471684	0.02687027	0.009960302
7	-0.03423954	-0.031426	0.02039217
10	-0.03725384	-0.03113289	0.01891327
12	-0.005405632	-0.03910277	0.03133423
13	-0.03851884	0.038542	0.000365755
14	-0.02207059	0.03106015	0.009628402
16	-0.02131568	0.02740285	0.007668
17	-0.03225324	0.03651002	0.004551062
19	-0.02089678	0.03018141	0.008730951
20	-0.01577667	-0.05201167	-0.03981953
21	-0.03298104	0.03718242	0.002670315
26	-0.00499221	-0.05379721	-0.03013688
27	-0.007774852	0.02453685	0.01703427
28	-0.02366932	0.03210646	0.007533882
29	0.002220658	0.01755951	0.0222258
30	-0.008558646	0.02160268	0.01485408
32	-0.03123044	0.0359196	0.004760768
33	-0.002457004	0.0150931	0.01584006
36	-0.01606878	-0.0520305	-0.04090924
37	-0.026679	0.03384569	0.005838128
38	-0.03301569	0.03641526	0.002929629
39	0.001335071	-0.05579934	-0.0267268
41	-0.02899757	0.03501667	0.005019332
43	-0.002192114	-0.05469876	-0.02890374
44	-0.007991841	-0.05452079	-0.03443027
45	0.000122855	-0.05483949	-0.02800839
47	-0.01525136	0.02523554	0.01283152
50	-0.01585745	-0.04926248	-0.0354027
51	-0.04421051	-0.05817104	-0.01326759
54	0.01505416	-0.05225585	-0.01420127
59	-0.02673015	0.03442829	0.006106066
64	-0.01421821	0.02576568	0.01220391
71	-0.01539269	-0.05211509	-0.03970882
74	-0.01495915	0.02641009	0.01261957
76	-0.03868801	-0.04352402	-0.0497501
77	-0.02338422	0.03070152	0.007687599
78	-0.07002623	-0.05221909	-0.02093921
79	-0.01724063	0.02847817	0.01244097
85	-0.03401601	0.0360208	0.001900634
86	-0.005931722	0.01665723	0.01546239
87	0.01841025	0.005398996	0.02752936
88	-0.02916694	-0.04858757	-0.04850207
89	-0.02954953	0.03317678	0.004209197
90	-0.025802	-0.04959363	-0.04757768
91	0.0100168	0.007150819	0.02209383
92	-0.02855206	0.0346554	0.006624161
96	-0.02774313	-0.04940543	-0.04677997
98	-0.03804618	0.03960855	0.000348765
100	-0.02678841	-0.04873691	-0.04736834

101	-0.03791046	0.03815397	-8.50E-05
107	-0.04067874	0.03931888	-0.000671054
109	-0.002866116	-0.05312126	-0.02980177
110	-0.0326237	0.03492853	0.001571162
114	0.03759081	-0.01162451	0.03713438
116	-0.01137374	0.02019846	0.01277787
117	-0.03278205	0.03611893	0.0031502
119	0.007562798	0.01348449	0.02555317
120	-0.02165749	0.0288201	0.008404752
123	-0.02731104	0.03365515	0.004673102
124	-0.04051552	0.03993622	-0.000910519
125	-0.03763423	0.03802576	0.000383601
126	0.3604962	0.007302757	0.001593934
127	-0.02741537	-0.0494845	-0.04756646
130	-0.02705983	-0.04987263	-0.04602252
131	-0.03625262	0.03707012	0.00115074
132	-0.03211531	0.03571649	0.002919779
136	-0.01064183	0.02432693	0.01634595
137	-0.03719145	0.03799149	0.001335781
138	-0.03518501	0.03817541	0.002692697
139	-0.03832641	0.03932942	-0.000474035
140	-0.0178982	0.02915306	0.01224065
141	-0.00859681	0.022185	0.01657161
149	-0.006392624	0.02400159	0.01702621
150	-0.04191278	0.04111145	-0.003317331
154	-0.0198634	0.03003133	0.00924166
156	-0.03473089	0.03495332	0.000660076
158	-0.041075	0.04044214	-0.002027919
161	-0.001873619	0.02043122	0.01848757
162	-0.02018061	0.02945914	0.008710454
166	-0.02320155	0.03246096	0.008223116
167	0.2949488	0.006946185	0.006633996
171	-0.03369286	0.04014914	0.002010747
172	-0.03884067	0.03918405	-0.00014204
173	-0.007546283	0.02242039	0.01573319
176	0.3474135	0.008397177	0.001160447
189	0.1433435	-0.06400331	-0.3031739
193	0.265822	-0.01382262	0.00902535
217	0.343524	0.0061475	0.001670636
219	0.195724	-0.02697771	0.02241014
225	-0.02711142	0.03414031	0.005895727
238	-0.01795593	-0.05158628	-0.04069377
239	0.3119844	0.005379935	0.008360796
243	-0.009122306	-0.03770218	0.03020156
246	0.08352907	-0.03957626	0.04004538
249	-0.02525274	0.03318012	0.007116607
252	-0.04127489	0.04043713	-0.001459343
263	-0.012601	-0.07864169	0.10266
265	-0.05588884	-0.08074756	0.1105057
266	-0.001632996	-0.001725051	0.01112865
268	-0.03495953	0.03309895	0.00027412
270	-0.03318577	-0.07649457	0.06908
272	-0.04172257	-0.07807688	0.1097543
273	-0.03087615	-0.082852	0.1081027

274	0.001606071	-0.07340068	0.09864818
276	-0.04532844	-0.08401243	0.1118804
278	-0.02348644	-0.08579282	0.09750047
280	-0.02727128	0.03325419	0.006196847
296	-0.02132793	0.04075289	0.005228211
297	0.000392189	0.01565972	0.02389062
299	-0.0186546	-0.03556889	0.02635953
300	-0.01047532	-0.05767409	0.04118689
301	0.02058858	-0.05251171	0.05353189
302	-0.007582879	-0.05275123	0.0534237
303	-0.03230772	0.03675152	0.003098798
314	-0.02457106	0.02725304	0.005418598
315	0.01428284	-0.04307615	0.03466308
324	-0.0173042	0.02944561	0.01041132
328	0.000519177	-0.04209551	0.03414676
332	0.008781774	0.01190923	0.0244555
334	-0.02993797	-0.03297232	0.02229911
335	-0.01187331	-0.05426197	-0.03679238
338	-0.01848723	0.02437947	0.007307517
339	-0.003409787	0.01761063	0.01917681
342	0.006691199	0.01372537	0.0245139
343	-0.0306033	0.03700462	0.004180157
344	-0.0270108	0.03525221	0.006366232
348	-0.003746486	-0.04099536	0.03254993
349	-0.02561723	0.03229977	0.006381006
353	0.2815823	0.005608561	0.008717716
354	0.2229905	-0.000530611	0.0113674
356	-0.01797567	0.02927512	0.01162471
358	-0.03136855	-0.03311856	0.02147454
360	-0.02127457	-0.05073889	-0.04463512
363	-0.01728313	-0.03660787	0.02821808
364	0.2435856	-0.000254961	0.01206529
366	-0.03987133	0.03952176	-0.000151324
367	-0.04187495	-0.02522165	0.02205501
368	-0.02463285	-0.05011929	-0.04646482
369	-0.03597576	-0.03164533	0.0204955
373	-0.01524926	-0.08052351	0.1045181
376	-0.006536389	-0.05505363	-0.0325893
377	-0.03784319	-0.03035915	0.01891667
379	-0.02977059	0.03333672	0.004595609
380	-0.02010871	-0.05076481	-0.04287576
386	0.3096348	0.009106378	0.00564991
389	0.007836792	-0.05706042	-0.02148587
390	-0.02499578	0.03160669	0.00733603
391	-0.0463498	-0.0795368	0.109986
394	-0.01664416	-0.03698585	0.02766837
396	-0.04081338	-0.08120384	0.1122367
403	-0.02231721	-0.05002131	-0.04318055
404	-0.01317995	-0.05189302	0.05177339
409	-0.02615366	-0.04879416	-0.04647823
418	-0.03254861	0.03055989	0.000873976
423	-0.02221038	-0.05107184	-0.04570088
424	-0.01778672	-0.05221514	-0.04155841
425	-0.02175821	-0.05031657	-0.04387976

427	-0.03419812	0.03684233	0.001988576
429	-0.04086707	0.0380805	-0.001142787
430	-0.05756495	0.0316489	0.006863468
431	-0.0220611	-0.05013843	-0.04376006
432	0.001764403	-0.05519532	-0.02584134
433	-0.02053101	-0.05077205	-0.04337824
434	-0.03033273	-0.04784374	-0.04955264
438	-0.02765943	-0.04973803	-0.04963189
440	0.01008159	-0.05557706	-0.01868682
444	-0.04610853	0.04116348	-0.004029137
449	-0.04596381	0.04138073	-0.003950088
452	-0.02435507	-0.05033297	-0.04512996
454	-0.04184351	0.03743946	-0.001450746
457	-0.02493109	-0.05040162	-0.04639369
459	0.239237	0.01078363	0.006372887
461	-0.02624122	-0.04935343	-0.0472484
464	-0.02875035	-0.04876851	-0.04853753
466	-0.01838866	-0.05115269	-0.04171299
467	-0.02776578	-0.04944165	-0.04770856
469	-0.02773447	-0.04959491	-0.05764467
470	-0.02904309	-0.04858716	-0.04847144
486	-0.03271321	0.0365002	0.002613259
487	-0.02986178	0.03539367	0.005649897
489	-0.02229359	0.03092937	0.008445284
491	-0.0286083	0.03549827	0.005125927
494	-0.007252257	-0.05335436	-0.03317951
497	-0.03285708	0.03671533	0.004523378
502	-0.04027182	0.04052107	-0.000663786
504	0.260876	-0.01349185	0.001975083
506	-0.03661007	0.03799979	0.00082465
510	0.005649093	-0.04202536	0.03653783
511	-0.03006926	-0.03317592	0.0221951
513	-0.03722219	-0.03094186	0.01862624
514	-0.03807562	-0.0304848	0.01765541
517	-0.03527791	0.03248639	-0.000229652
520	-0.007441262	-0.05357097	-0.03395338
522	0.3135348	0.007617704	0.006784224
523	0.2250985	-0.01584105	0.01054844
526	-0.04381245	0.04149401	-0.002948693
530	-0.03155595	0.03197118	0.001737534
531	-0.02753505	-0.04908207	-0.04839896
533	-0.02134859	-0.05014928	-0.04366154
536	-0.02740957	-0.04921533	-0.04761115
537	-0.03293108	0.03478751	0.004076194
540	-0.03647808	0.03293092	0.000389882
541	-0.03210577	0.03504821	0.004097848
542	-0.01733322	0.02763402	0.01216452
543	-0.03421773	0.0352974	0.002333241
545	-0.03022313	0.03566865	0.00354519
547	-0.01841563	0.02863601	0.0101051
550	-0.03710994	0.03591351	0.001148822
553	-0.02345793	-0.04926149	-0.04870546
556	-0.01734223	-0.05193765	-0.04159019
559	-0.01124832	-0.05161133	-0.03579888

566	-0.005460243	0.01958446	0.01673633
568	-0.03327248	0.03732435	0.002206033
570	-0.006961535	0.0197962	0.01693823
572	0.008033432	0.01180223	0.02195854
574	-0.004252202	0.01705932	0.01699314
577	-0.02678016	0.03079644	0.006418302
578	0.00874025	0.008909502	0.02434315
582	-0.02781565	0.03270748	0.006829523
583	0.006251986	-0.05532552	-0.02248053
584	-0.01055651	0.02214582	0.01547296
585	-0.04669104	0.04358724	-0.005429165
587	-0.02690472	0.03188144	0.006383008
588	-0.0306366	0.03263653	0.004450965
590	-0.04069714	0.03909661	-0.0016809
591	-0.04003829	0.04103045	-0.001198396
592	0.003398989	0.01250413	0.02248318
593	-0.01507802	0.02461076	0.01195545
595	-0.04154096	0.0379748	-0.001569197
596	-0.009477662	0.02194097	0.01595409
598	-0.03623689	0.03830459	0.001647345
599	0.007095494	0.00887617	0.02157535
600	-0.006581995	0.02224112	0.01622614
603	-0.01440874	0.02660056	0.0112886
604	-0.02433831	0.03147704	0.00740573
605	-0.00374848	0.01495812	0.01565401
607	-0.01201968	0.0236293	0.01214226
609	-0.007383813	0.0175602	0.01463242
614	-0.03048994	0.03534194	0.005251731
615	-0.03155328	0.03593242	0.004004195
616	-0.0277518	0.03642185	0.00483795
617	-0.03974241	0.03896445	-0.000228614
618	0.04663247	-0.0190984	0.04235293
619	-0.0344351	0.03683471	0.000583476
627	-0.03786412	0.03751005	0.000839847
630	-0.01997345	0.01442047	0.004092575
633	-0.02410781	0.02560573	0.005326476
634	-0.02420338	-0.04993722	-0.04561961
635	-0.02753002	0.0334259	0.004506002
636	-0.04304184	0.03556391	-0.004211435
637	-0.0456729	0.040011	-0.004020634
641	-0.04032922	0.0333795	-0.003223325
642	-0.04514082	0.04137708	-0.003981785
643	-0.02988742	-0.04926802	-0.04806751
651	-0.03135767	0.03066513	0.002256069
652	-0.02970808	0.0340769	0.004516528
654	-0.03054128	0.03473599	0.003922279
658	0.1959039	-0.02502756	0.02257753
659	0.2519944	-0.01397722	0.005775191
660	0.2925307	0.006216676	0.006656262
661	0.198851	-0.02706336	0.02153422
664	0.08026186	-0.04275036	0.04854714
669	0.303512	0.007468524	0.003233941
670	-0.03286272	0.03707715	0.002355624
672	-0.0204379	-0.05050186	-0.04221584

673	0.07841786	-0.04153475	0.04841083
674	-0.0444137	-0.02875208	0.01640324
675	-0.04515026	-0.02865293	0.01545245
676	-0.04214561	-0.02917087	0.01646312
680	-0.04541995	0.04120365	-0.004407666
684	-0.03760211	-0.0295186	0.01642277
685	-0.02203346	-0.05028071	-0.04300131
687	-0.01316877	-0.07917601	0.1035414
688	0.2747302	-0.009528782	0.01209346
693	-0.03714632	0.03530229	0.000403456
694	-0.0327546	0.03637824	0.003680747
696	-0.03725409	0.03738844	0.000522724
697	-0.03560497	0.03895078	0.001762207
698	-0.03151181	0.03682159	0.003412309
701	-0.03481513	0.03617468	0.002265161
702	-0.03876389	0.03657479	-0.001721476
703	-0.03616479	0.05086194	-0.00386477
705	-0.04560175	0.04151011	-0.004055435
707	-0.02731065	-0.04866386	-0.04817728
708	-0.04351968	0.04088109	-0.003250039
709	-0.04601992	0.04227878	-0.004302788
711	-0.04056365	0.03974477	-0.002023492
713	0.3137059	0.01217421	0.003162078
715	-0.04451539	0.0416938	-0.003739684
716	0.3625383	0.01377377	-0.004613305
717	-0.03273939	-0.03181807	0.02043739
718	-0.035447	-0.03348421	0.01948678
720	-0.0378516	0.03868592	-0.000339678
722	-0.0285888	-0.03305896	0.02199933
723	-0.04073625	-0.02894139	0.01786783
724	-0.04229827	-0.03191634	0.01425711
725	-0.0267103	-0.04927423	-0.0457125
726	-0.0393071	-0.03057295	0.01798196
727	-0.02856963	-0.04824619	-0.04712658
729	-0.01098067	-0.03726713	0.04070788
731	-0.03204045	0.03093295	0.000377412
732	-0.03341259	0.03173951	0.000622464
733	-0.02173763	0.03032571	0.008873087
735	-0.0301697	-0.0329238	0.02226234
736	-0.03532506	0.03786753	0.001585515
739	-0.03024336	-0.04903731	-0.04775191
742	-0.03645904	0.03747788	0.000553346
744	-0.03308956	0.03624683	0.002890389
745	-0.03343167	-0.05016971	-0.0508086
747	-0.03481853	0.03858695	0.001531808
750	-0.02356529	-0.03478734	0.02343503
752	-0.0273686	0.03466374	0.006912698
753	-0.02014604	0.02686228	0.009420495
754	-0.01778095	-0.03846922	0.03822214
757	-0.006964035	-0.03918043	0.03020983
760	-0.0316562	0.03610473	0.003492828
762	0.3335657	0.003514547	0.000732208
763	-0.01427624	-0.0511908	-0.03724012
767	0.3128255	0.008203532	0.00524089

769	-0.02238682	0.031754	0.0100507
770	0.3041768	0.001215272	0.003442606
771	-0.03626493	-0.03117568	0.01882899
772	-0.02601381	0.03326453	0.006176731
774	-0.02494816	0.03430527	0.006867379
775	-0.03356675	-0.03252473	0.02218427
776	-0.01370064	0.02645121	0.01375667
777	-0.01562424	-0.05189418	-0.03772476
778	-0.02980113	-0.03373328	0.02248968
779	-0.0179494	0.02915005	0.01032483
780	-0.02563019	0.0338792	0.006650796
781	-0.03636576	0.03930899	0.000605408
783	-0.02850355	0.03551007	0.00498894
785	0.2945877	0.000648804	0.007310936
787	0.3355265	0.005317825	0.002299309
788	-0.03041286	0.03896433	0.002710719
789	-0.01456717	-0.05387356	-0.03924547
793	-0.0277209	0.03449862	0.005916476
800	-0.01152834	-0.05238925	-0.03738696
802	0.2754906	-0.01362632	0.004974513
804	-0.02572981	0.02833783	0.004482017
805	0.05217566	-0.02399688	0.04330838
806	0.1606538	0.3468447	-0.1248018
807	-0.04266641	-0.02518356	0.01719488
808	-0.01787181	0.02326605	0.008550738
809	-0.02104423	-0.05046771	-0.04241698
810	-0.01284132	-0.05278735	-0.0369328
811	-0.0134386	-0.05288894	-0.0367293
812	-0.01745845	-0.05143214	-0.04084192
813	-0.02828442	0.03514728	0.004331266
814	0.01302525	0.01183222	0.02751409
815	-0.02653427	-0.03418933	0.02286389
816	-0.02741792	-0.03388968	0.02337522
821	-0.02262456	-0.05201587	-0.04326422
822	-0.01833988	-0.05010777	-0.04234396
823	-0.02098936	-0.05090561	-0.04492109
824	-0.03146294	-0.03265838	0.02110553
827	-0.0108951	-0.05288688	-0.03667355
831	-0.02996728	-0.03481846	0.02268515
832	-0.02491573	-0.05001683	-0.04573844
833	-0.02639413	-0.05152976	-0.04585056
834	-0.03401424	0.03645546	0.00263904
835	-0.03427175	0.036052	0.00171697
837	-0.01921635	-0.05110828	-0.04091295
838	-0.01088035	-0.05191315	-0.03456421
840	0.07010846	-0.04549534	0.04800254
842	0.101544	0.3027296	-0.1086104
843	-0.03772574	0.03876468	-0.000127047
845	-0.04104001	0.03829991	-0.001272796
846	-0.007045335	-0.05305888	-0.03342288
847	-0.02563709	0.03048031	0.006079886
871	-0.03056935	0.02995714	0.001196601
872	-0.02174644	-0.05084379	-0.04349218
873	-0.02251504	-0.05076971	-0.04323009

874	-0.02806219	0.03459157	0.005428022
875	-0.0355574	0.03837039	0.001641124
876	-0.04373489	0.04015245	-0.002333622
877	-0.0326239	0.03791655	0.003443213
878	-0.04557428	0.04114586	-0.004021003
879	-0.02898666	0.02861571	0.002596491
893	-0.04532564	0.04103336	-0.003930002
900	0.3256594	0.009140757	0.004363542
901	-0.03907242	0.03287794	-0.002850247
902	-0.04312621	0.04218286	-0.004601734
903	0.02181915	-0.05722186	-0.009313113
905	-0.01752152	-0.0522496	-0.04053021
907	-0.02269687	-0.05031509	-0.04582537
908	-0.0359696	0.03864036	0.001958788
909	-0.03109282	0.03136307	0.002334395
911	-0.02756139	0.02957396	0.003302971
950	-0.03069982	0.02970748	0.002103021
951	-0.009155965	-0.05287329	-0.03348038
952	-0.01786286	-0.05211286	-0.04142749
953	-0.02514014	0.03351999	0.007364428
955	-0.01755768	-0.05128171	-0.04516505
956	-0.02167369	-0.05096791	-0.04783293
958	-0.02683305	0.03305221	0.006289356
959	-0.03120778	0.03640036	0.004074912
960	-0.04258505	0.04137002	-0.003162678
962	-0.04161794	0.04169291	-0.002526043
963	-0.04453179	0.04082018	-0.003500275
964	-0.04206249	0.04048754	-0.002751883
965	-0.03827856	0.03391775	-0.002493336
966	-0.02905686	0.0342182	0.001135981
1000	-0.02178357	-0.05111719	-0.04465502
C1	-0.03306864	-0.03224527	0.02057606
C2	-0.01747971	-0.03616587	0.02755607
C3	-0.03096341	0.03595748	0.004094081
C4	-0.004355531	-0.05485881	-0.03143992
C5	-0.03438258	0.03718092	0.002521741
G	0.02306741	0.000712435	0.03114293
K	-0.03292104	0.0367014	0.003400073
L	-0.03595115	0.03593434	0.001557385
M	-0.01625093	-0.03667559	0.03038043

Supplementary Material 1.5. STRUCTURE-inferred values.

ID	Cluster 1	Cluster 2	Cluster 3
1	0.551	0.171	0.278
2	0.548	0.169	0.283
3	0.540	0.182	0.277
4	0.198	0.662	0.139
7	0.014	0.974	0.013
10	0.009	0.979	0.012
12	0.052	0.925	0.023
13	0.549	0.169	0.282
14	0.552	0.167	0.281
16	0.555	0.162	0.283
17	0.549	0.168	0.283
19	0.548	0.172	0.280
20	0.540	0.195	0.265
21	0.549	0.170	0.281
26	0.551	0.183	0.266
27	0.558	0.164	0.278
28	0.547	0.173	0.280
29	0.554	0.170	0.277
30	0.558	0.159	0.282
32	0.555	0.163	0.282
33	0.548	0.177	0.274
36	0.533	0.198	0.270
37	0.556	0.162	0.282
38	0.555	0.168	0.277
39	0.540	0.193	0.266
41	0.548	0.171	0.281
43	0.038	0.938	0.024
44	0.546	0.187	0.267
45	0.536	0.201	0.262
47	0.554	0.169	0.278
50	0.472	0.292	0.237
51	0.016	0.947	0.037
54	0.543	0.190	0.267
59	0.550	0.167	0.284
64	0.561	0.157	0.282
71	0.533	0.202	0.266
74	0.553	0.168	0.278
76	0.443	0.320	0.237
77	0.551	0.164	0.286
78	0.008	0.980	0.012
79	0.559	0.158	0.283
85	0.551	0.169	0.280
86	0.521	0.210	0.269
87	0.553	0.173	0.274
88	0.538	0.191	0.271
89	0.542	0.181	0.277
90	0.531	0.205	0.264
91	0.560	0.164	0.276
92	0.556	0.158	0.286
96	0.485	0.262	0.253
98	0.539	0.175	0.285
100	0.542	0.192	0.266

101	0.549	0.173	0.278
107	0.550	0.163	0.287
109	0.530	0.216	0.255
110	0.536	0.190	0.274
114	0.568	0.148	0.284
116	0.555	0.167	0.278
117	0.550	0.168	0.281
119	0.562	0.152	0.285
120	0.552	0.162	0.287
123	0.554	0.164	0.281
124	0.551	0.168	0.281
125	0.548	0.170	0.282
126	0.030	0.039	0.931
127	0.536	0.200	0.264
130	0.532	0.201	0.267
131	0.549	0.172	0.279
132	0.546	0.175	0.279
136	0.554	0.165	0.281
137	0.553	0.167	0.280
138	0.553	0.169	0.279
139	0.551	0.168	0.280
140	0.549	0.174	0.277
141	0.556	0.166	0.278
149	0.551	0.166	0.283
150	0.541	0.178	0.281
154	0.547	0.170	0.283
156	0.548	0.172	0.280
158	0.540	0.178	0.282
161	0.553	0.160	0.287
162	0.556	0.164	0.280
166	0.558	0.158	0.283
167	0.200	0.091	0.709
171	0.538	0.171	0.291
172	0.545	0.171	0.283
173	0.553	0.164	0.283
176	0.044	0.032	0.924
189	0.002	0.004	0.994
193	0.202	0.083	0.714
217	0.034	0.036	0.930
219	0.313	0.103	0.584
225	0.557	0.156	0.288
238	0.531	0.208	0.262
239	0.162	0.089	0.749
243	0.032	0.942	0.026
246	0.487	0.180	0.333
249	0.549	0.171	0.280
252	0.543	0.175	0.282
263	0.394	0.323	0.283
265	0.377	0.332	0.290
266	0.426	0.349	0.225
268	0.548	0.171	0.282
270	0.313	0.429	0.258
272	0.378	0.334	0.288
273	0.388	0.326	0.286

274	0.399	0.327	0.273
276	0.373	0.335	0.291
278	0.362	0.361	0.276
280	0.552	0.167	0.281
296	0.535	0.149	0.316
297	0.550	0.169	0.280
299	0.019	0.964	0.017
300	0.395	0.346	0.258
301	0.422	0.315	0.263
302	0.385	0.335	0.280
303	0.544	0.178	0.278
314	0.551	0.166	0.283
315	0.433	0.284	0.284
324	0.554	0.167	0.279
328	0.073	0.892	0.035
332	0.554	0.167	0.279
334	0.010	0.980	0.010
335	0.538	0.200	0.262
338	0.550	0.168	0.282
339	0.551	0.171	0.278
342	0.561	0.156	0.283
343	0.555	0.163	0.283
344	0.551	0.170	0.279
348	0.050	0.918	0.032
349	0.554	0.162	0.284
353	0.211	0.076	0.714
354	0.229	0.093	0.678
356	0.549	0.174	0.276
358	0.012	0.980	0.007
360	0.538	0.191	0.271
363	0.029	0.953	0.018
364	0.245	0.083	0.672
366	0.541	0.181	0.278
367	0.008	0.983	0.009
368	0.536	0.195	0.269
369	0.009	0.983	0.009
373	0.399	0.319	0.282
376	0.538	0.201	0.261
377	0.009	0.984	0.007
379	0.545	0.178	0.277
380	0.543	0.192	0.265
386	0.196	0.094	0.710
389	0.541	0.193	0.265
390	0.549	0.171	0.279
391	0.377	0.334	0.288
394	0.023	0.960	0.017
396	0.386	0.322	0.292
403	0.527	0.210	0.262
404	0.385	0.338	0.278
409	0.537	0.193	0.271
418	0.546	0.176	0.278
423	0.537	0.200	0.263
424	0.543	0.194	0.263
425	0.544	0.190	0.266

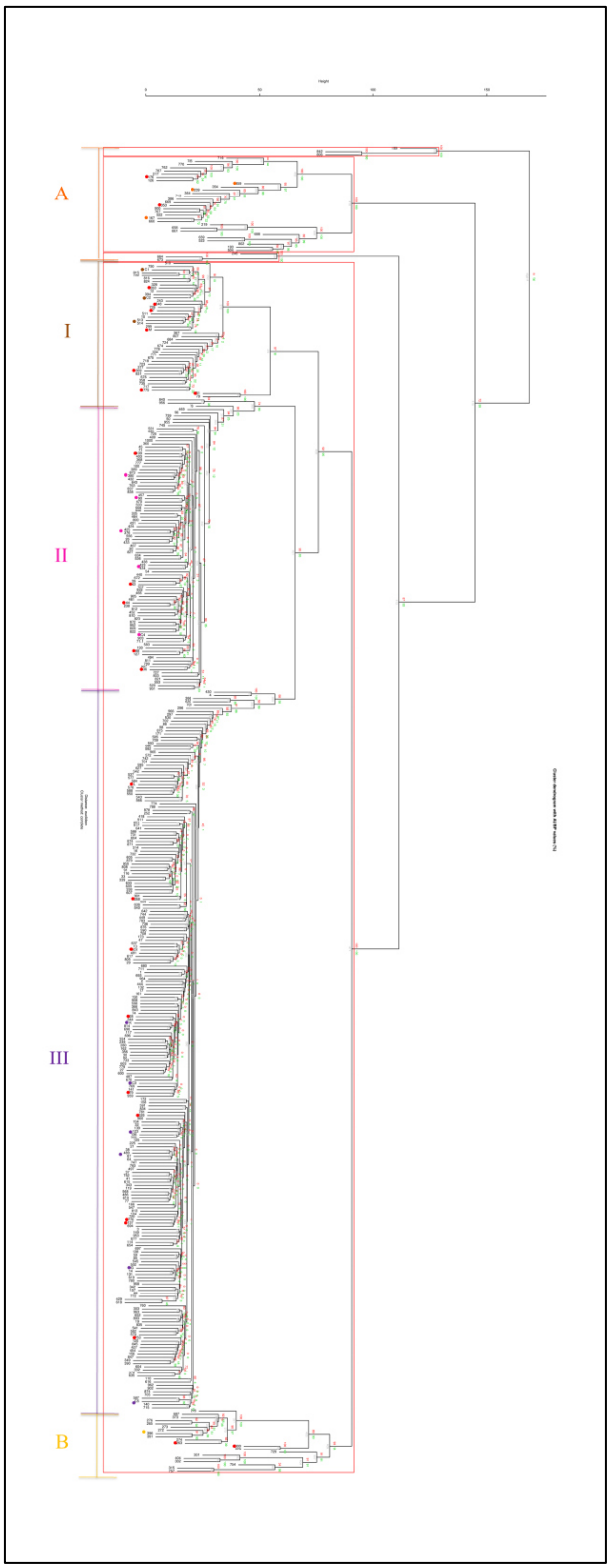
427	0.549	0.169	0.282
429	0.547	0.173	0.280
430	0.286	0.549	0.165
431	0.540	0.199	0.262
432	0.537	0.198	0.264
433	0.541	0.194	0.265
434	0.541	0.190	0.269
438	0.531	0.203	0.266
440	0.538	0.205	0.257
444	0.541	0.180	0.279
449	0.541	0.180	0.279
452	0.533	0.203	0.264
454	0.546	0.173	0.280
457	0.531	0.202	0.267
459	0.209	0.085	0.706
461	0.537	0.191	0.272
464	0.539	0.194	0.267
466	0.527	0.212	0.261
467	0.543	0.190	0.267
469	0.519	0.193	0.288
470	0.533	0.203	0.264
486	0.555	0.158	0.286
487	0.553	0.165	0.282
489	0.557	0.162	0.282
491	0.550	0.168	0.282
494	0.523	0.215	0.262
497	0.547	0.172	0.280
502	0.546	0.174	0.280
504	0.184	0.084	0.732
506	0.552	0.164	0.284
510	0.094	0.853	0.053
511	0.011	0.976	0.013
513	0.012	0.978	0.011
514	0.008	0.980	0.012
517	0.540	0.179	0.281
520	0.525	0.209	0.265
522	0.202	0.084	0.714
523	0.220	0.064	0.716
526	0.553	0.162	0.284
530	0.514	0.213	0.273
531	0.528	0.208	0.265
533	0.532	0.206	0.262
536	0.534	0.199	0.266
537	0.547	0.173	0.279
540	0.547	0.176	0.277
541	0.546	0.176	0.278
542	0.552	0.166	0.281
543	0.541	0.182	0.277
545	0.552	0.163	0.285
547	0.559	0.158	0.284
550	0.547	0.177	0.277
553	0.531	0.194	0.275
556	0.539	0.196	0.265
559	0.536	0.202	0.263

566	0.553	0.161	0.286
568	0.551	0.171	0.278
570	0.554	0.170	0.277
572	0.547	0.176	0.277
574	0.558	0.158	0.284
577	0.544	0.178	0.278
578	0.561	0.159	0.281
582	0.547	0.175	0.278
583	0.524	0.221	0.255
584	0.546	0.177	0.277
585	0.537	0.180	0.284
587	0.544	0.181	0.275
588	0.547	0.175	0.278
590	0.540	0.176	0.284
591	0.533	0.188	0.278
592	0.534	0.198	0.267
593	0.537	0.189	0.274
595	0.539	0.182	0.279
596	0.539	0.188	0.273
598	0.550	0.168	0.283
599	0.551	0.172	0.277
600	0.559	0.162	0.279
603	0.552	0.171	0.277
604	0.543	0.182	0.274
605	0.560	0.158	0.282
607	0.554	0.165	0.280
609	0.554	0.168	0.278
614	0.550	0.167	0.282
615	0.556	0.162	0.282
616	0.550	0.164	0.286
617	0.551	0.168	0.281
618	0.564	0.163	0.273
619	0.553	0.168	0.279
627	0.543	0.179	0.277
630	0.473	0.274	0.253
633	0.553	0.165	0.282
634	0.537	0.197	0.267
635	0.554	0.165	0.282
636	0.540	0.184	0.275
637	0.551	0.167	0.283
641	0.551	0.171	0.278
642	0.536	0.185	0.279
643	0.535	0.200	0.265
651	0.549	0.170	0.281
652	0.545	0.175	0.280
654	0.544	0.173	0.283
658	0.316	0.088	0.597
659	0.155	0.072	0.773
660	0.202	0.085	0.713
661	0.307	0.106	0.587
664	0.466	0.248	0.286
669	0.196	0.077	0.727
670	0.557	0.157	0.286
672	0.536	0.198	0.266

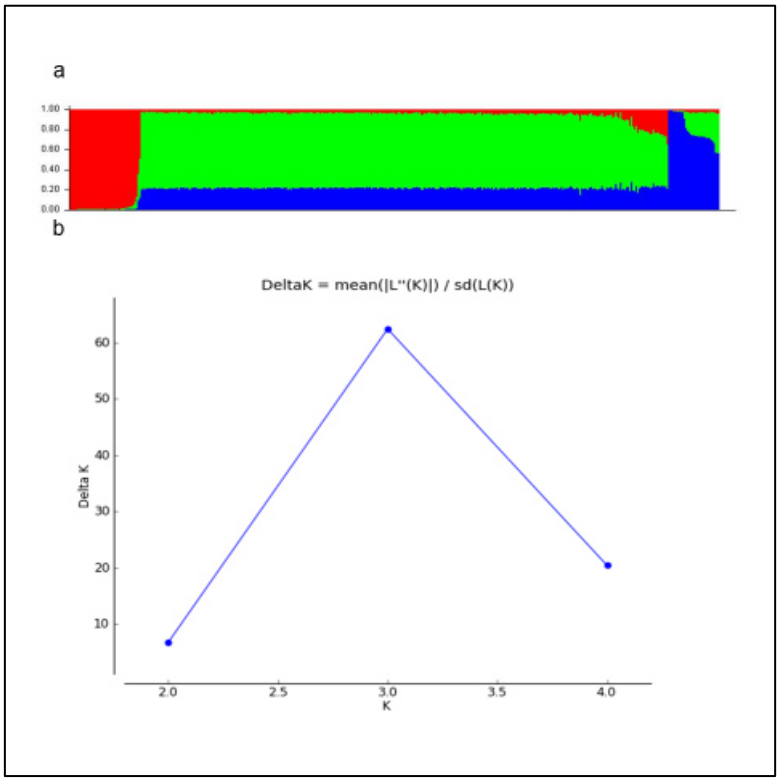
673	0.490	0.216	0.294
674	0.009	0.981	0.011
675	0.010	0.980	0.011
676	0.008	0.985	0.007
680	0.539	0.180	0.281
684	0.013	0.973	0.013
685	0.522	0.215	0.263
687	0.397	0.329	0.274
688	0.214	0.082	0.704
693	0.549	0.173	0.278
694	0.551	0.170	0.279
696	0.547	0.170	0.283
697	0.549	0.171	0.281
698	0.553	0.163	0.284
701	0.547	0.168	0.285
702	0.531	0.183	0.286
703	0.509	0.167	0.324
705	0.540	0.180	0.279
707	0.547	0.179	0.274
708	0.546	0.173	0.281
709	0.540	0.171	0.289
711	0.542	0.178	0.280
713	0.187	0.082	0.732
715	0.543	0.170	0.288
716	0.039	0.028	0.933
717	0.008	0.982	0.011
718	0.009	0.979	0.012
720	0.549	0.169	0.282
722	0.014	0.972	0.014
723	0.008	0.981	0.011
724	0.008	0.980	0.012
725	0.526	0.214	0.260
726	0.010	0.979	0.010
727	0.517	0.221	0.262
729	0.375	0.320	0.305
731	0.554	0.164	0.282
732	0.545	0.174	0.281
733	0.561	0.157	0.282
735	0.010	0.979	0.011
736	0.543	0.176	0.281
739	0.484	0.266	0.250
742	0.543	0.174	0.283
744	0.546	0.172	0.283
745	0.511	0.232	0.257
747	0.550	0.169	0.282
750	0.018	0.970	0.011
752	0.554	0.165	0.281
753	0.553	0.167	0.280
754	0.401	0.320	0.279
757	0.409	0.302	0.289
760	0.545	0.173	0.282
762	0.031	0.030	0.940
763	0.536	0.202	0.262
767	0.194	0.081	0.725

769	0.551	0.168	0.281
770	0.042	0.043	0.915
771	0.010	0.982	0.008
772	0.552	0.170	0.278
774	0.544	0.170	0.286
775	0.009	0.979	0.012
776	0.560	0.157	0.284
777	0.539	0.197	0.264
778	0.012	0.976	0.013
779	0.553	0.168	0.279
780	0.553	0.166	0.282
781	0.543	0.168	0.289
783	0.556	0.156	0.288
785	0.084	0.058	0.858
787	0.043	0.038	0.919
788	0.548	0.155	0.296
789	0.538	0.186	0.276
793	0.551	0.169	0.281
800	0.535	0.201	0.263
802	0.142	0.065	0.793
804	0.541	0.179	0.280
805	0.576	0.144	0.280
806	0.001	0.002	0.997
807	0.010	0.980	0.010
808	0.552	0.172	0.276
809	0.538	0.195	0.266
810	0.535	0.201	0.265
811	0.523	0.220	0.256
812	0.540	0.193	0.267
813	0.551	0.165	0.284
814	0.561	0.156	0.283
815	0.013	0.977	0.010
816	0.015	0.975	0.010
821	0.519	0.229	0.252
822	0.530	0.208	0.262
823	0.538	0.198	0.264
824	0.014	0.976	0.010
827	0.538	0.202	0.260
831	0.007	0.983	0.009
832	0.535	0.199	0.266
833	0.514	0.226	0.260
834	0.540	0.182	0.277
835	0.550	0.168	0.282
837	0.537	0.197	0.266
838	0.536	0.194	0.270
840	0.521	0.227	0.252
842	0.002	0.004	0.994
843	0.550	0.167	0.283
845	0.551	0.166	0.283
846	0.535	0.203	0.262
847	0.559	0.160	0.282
871	0.554	0.166	0.280
872	0.539	0.193	0.269
873	0.538	0.197	0.265

874	0.547	0.167	0.286
875	0.534	0.183	0.283
876	0.544	0.178	0.279
877	0.552	0.167	0.281
878	0.555	0.158	0.287
879	0.554	0.162	0.284
893	0.546	0.173	0.281
900	0.199	0.075	0.726
901	0.544	0.175	0.281
902	0.546	0.167	0.287
903	0.544	0.197	0.258
905	0.538	0.198	0.264
907	0.535	0.197	0.269
908	0.550	0.169	0.281
909	0.544	0.175	0.281
911	0.544	0.174	0.282
950	0.550	0.167	0.284
951	0.516	0.229	0.255
952	0.535	0.199	0.267
953	0.547	0.177	0.277
955	0.523	0.205	0.273
956	0.523	0.200	0.277
958	0.546	0.176	0.277
959	0.556	0.160	0.284
960	0.550	0.166	0.284
962	0.542	0.173	0.285
963	0.548	0.171	0.281
964	0.546	0.171	0.282
965	0.539	0.180	0.281
966	0.531	0.175	0.294
1000	0.530	0.197	0.273
C1	0.009	0.981	0.010
C2	0.017	0.968	0.014
C3	0.547	0.171	0.282
C4	0.529	0.211	0.260
C5	0.543	0.175	0.283
G	0.560	0.160	0.280
K	0.554	0.162	0.284
L	0.545	0.177	0.278
M	0.016	0.969	0.014



Supplementary Figure 1.1. Garlic dendrogram. Phylogenetic tree, with approximately unbiased (AU; red)/Bootstrap Probability (BP; green) percentage values and Euclidean distances, generated by complete-linkage method, to ascertain germplasm diversity. Cluster I includes C1 and C2 (Chinese varieties); Cluster II has C4 (Spanish White variety); and Cluster III shows C3 to C5 (Spanish Purple and Brazilian varieties). Samples C1 to C5, and others described in the text, are highlighted with colored dots. I corresponds to cluster II in STRUCTURE analysis, whereas II and III are equivalent to cluster I; and A and B correspond to cluster III using such software analysis. This figure can be obtained with higher resolution from <https://www.frontiersin.org/articles/10.3389/fgene.2017.00098/full#supplementary-material> (also included in the attached CD, available from Universidad de Córdoba).



Supplementary Figure 1.2. Garlic genetic structure. STRUCTURE software was used to analyze the studied garlic germplasm. (a) Diagram showing the three calculated clusters (K D 3); and (b) 1K values.

CHAPTER 2. Potential of DArTseq to identify polymorphic genes of interest in the absence of a reference genome

2.1. Abstract

In the previous chapter, DArTseq analyses generated 33,423 uncharacterized-polymorphic sequences. In this chapter, sequences were analyzed by Basic Local-Alignment Search Tool (BLAST). This way, 1,082 sequences were characterized from 110 different species, including *Allium sativum* and other *Allium* genera. 142 sequences were identified after filtering according to "identity" and "e-value" scores. Repeated sequences from different species were removed as well. From those sequences, 120 encoded proteins Then, Gene Ontology (GO) enrichment was performed and metabolic pathways were analyzed. This resulted in a total of 559 GO terms found, being 188 for Biological Process (BP), 122 for Cellular Component (CC), and 245 for Molecular Function (MF). Regarding metabolic pathways, 11 were detected. Among them, some related to lipid and carbohydrate metabolism, hormone signaling, and TriCarboxylic Acid (TCA) cycle. To do these analyses, since the garlic genome has not been sequenced, *Arabidopsis thaliana* (L.) Heynh GenBank entries were used. This lead to loss of some accessions that did not have correspondence in such genome. Nevertheless, data described in this chapter should be useful for further genetic and genomic studies in garlic. For instance, development of polymorphic molecular markers and identification of genes of interest with polymorphism. These include the ones encoding enzymes with industrial interest, as well as those involved in adaptation and defense against biotic or abiotic stresses.

2.2. Introduction

2.2.1. Sequence analyses

In the previous chapter, DArTseq analyses were performed in order to assess genetic diversity and structure of a garlic germplasm bank. To reduce complexity and perform GBS, thousands of polymorphic-Silico DArT markers were generated. Each one was a short sequence of approximately 69 bp. One of the advantages of DArTseq technique is that it is able to generate sequences without previous genomic knowledge of the species (Cruz et al., 2013). Conversely, sequences obtained this way are quite short, being therefore difficult to assemble into contigs, scaffolds or full genomes. This is particularly relevant for a species like garlic, without reference genome (Garavito et al., 2016). In any case, the purpose of DArTseq was not to sequence genomes, but to genotype accessions in an effective way.

On the other hand, there are fortunately a great panoply of tools to work *in silico,* even with non-model species, and in the absence of reference genomes. It is well known that nucleic acid and peptide sequences may share identities across species due to the evolution process (Horan et al., 2008). This makes possible finding related sequences to available ones, like DArTseq reads. This can be accomplished comparing available query sequences against databases, such as the ones at National Center for Biotechnology Information (NCBI) <https://www.ncbi.nlm.nih.gov> and Universal Protein Resource (UniProt) <http://www.uniprot.org>. Such tool allows to modify search parameters, depending on source query and search goal (Altschul et al., 1990). Moreover, classic First-Generation Sequencing (FGS), and mostly the higher throughput SGS and recently the TGS are generating huge amounts of sequencing data from many species, as previously reviewed (Dorado, Gálvez, et al., 2015). This has exponentially increased the probability to find identities to a particular query sequence across available ones in databases in recent years (Xianjun et al., 2014).

This may be quite relevant, allowing to assign identity and function to otherwise unknown genes of species without reference genome (Horan et al., 2008), as in the case of garlic. Then, once sequence identities are found, it is possible to ascertain putative

functions and using GO. The latter is used to describe annotated attributes of gene products. Annotations are performed in three big and non-overlapping domains in molecular biology: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). Ontologies provide conceptualizations of domains, which are useful to share data for many different purposes in the current frame of "-omic" sciences and "big data". Relationships among GO terms are done by using "is-a" and "part-of" structured-vocabulary terms. BP shows general processes; for instance, response to stresses or photosynthesis. CC describes subcellular locations, where processes are taking place; for instance, nucleus or cytoplasm. Finally, MF describes molecular-level activities, but without specifying where or when they are carried out; i.e., metabolic or binding. Annotated GO terms usually are the result of collaborations between different research projects. To ensure quality, annotations ought to be linked to a source. For example, literature, computational analyses, or databases. To achieve high-quality annotations, both, literature and experimental support should be provided (Harris et al., 2004).

Due to the huge amount of accumulated GO data in its three domains, GO slims have been developed. They are summarized versions of GO that allow to have a generalized and broad view of the three ontologies. They are extremely useful in order to report GO annotation results. The first time GO slims were used was for annotation of *Drosophila melanogaster* Meigen genome (Adams et al., 2000). Summing up, GO slims make easier and quicker to perform an overall study of GO-term distributions (Harris et al., 2004).

2.2.2. Objectives

The main objective of this chapter is to ascertain the potential of DArTseq to identify genes of interest, in the absence of reference genome. As a proof of concept, it has been applied to garlic, which reproduces asexually and has a virtually unknown, large and expected complex genome. The specific objectives are: i) analyze garlic sequences obtained by DArTseq; ii) detect identities with other sequenced species; iii) perform enrichment analyses with GO terms; and iv) find metabolic pathways related to such genes.

2.3. Materials and methods

A total of 33,423 sequences obtained from DArTseq (SilicoDArT markers) were used in this chapter. Polymorphism-Information Content (PIC) values were included for each marker. First, a Fast-Alignment Sequence Tools (FAST)-All (FASTA) format file was created and analyzed with BLAST (NCBI) version 2.6.0 (Altschul et al., 1990). Specifically, with Standard Nucleotide BLAST (BLASTn) tool. Algorithm parameters were left as default. In short, only first 100 results were shown, the search was adjusted for short sequences, the expect threshold was 10 and minimum length was set as 28. Match score was 1 and mismatch score –2, gap costs were linear. Finally, low-complexity regions were filtered. Only sequences with "e-value" (that is, number of hits by chance) scores lower than 10^{-4} and identity higher than 80% were chosen. In order to perform BLASTx analyses, those that had a hit were searched against in UniProt database (The UniProt Consortium, 2016) for translated nucleotides, and to find the corresponding UniProt protein codes.

Afterwards, GO terms were searched in UniProt database. In addition, to summarize GO results, graphs were generated with REViGO tool (Supek et al., 2011) and Blast2GO software version 4.0 (Conesa et al., 2005). For GO-slim graph, Protein ANalysis THrough Evolutionary Relationships (PANTHER) classification system (Mi et al., 2013) was employed. Finally, the latter was also used, in order to find the metabolic pathways in which genes were involved. Input codes must belong to only one species for searching metabolic pathways. Therefore, UniProt codes from different species were changed to their corresponding homologous in just one species. The chosen species was *A. thaliana*, as it has huge genetic resources and a great number of UniProt entries. They have been manually curated and reviewed in SWISS-PROT (Horan et al., 2008). Input data was a UniProt ID list, using "Functional classification viewed in gene list", "Functional classification viewed in pie chart", and "Statistical overrepresentation test" with default settings for *A. thaliana* database.

2.4. Results

2.4.1. Sequence information and BLAST search

A total of 1,082 polymorphic sequences out of the 33,423 DArTseq reads or SilicoDArT markers (3.24%) had a hit in BLASTn database. First, samples were filtered according to their ID. For each repeated entry, hits were filtered first by identity and second by "e-value" score. Then, in the case of repeated entries, only the one with highest "identity" (that is, percentage of identity with the blasted sequence) and "e-value" scores were selected. After filtering, 142 GenBank entries were left (Supplementary Material 2.1). Maximum length was 69 bp, as delivered by this technique. The shortest sequences with homologies in the database had 32 bp. Average length was 65 bp. In relation to the 1,082 sequences with BLASTn hits, four exhibited identities with *A. sativum* entries. Specifically, 9341999_SilicoDArT marker was linked to "*Allium sativum* chitinase mRNA, 3' end" with Gene Identification (GI) 166342; 9323004_SilicoDArT to "*Allium sativum* phytochelatins synthase (*pcs1*) mRNA, complete cds" (GI 27448223); 9322412_SilicoDArT to "*Allium sativum* chloroplast cysteine synthase GCS2 (*gcs2*) mRNA, complete cds; nuclear gene for chloroplast product" (GI 59799342); and 9343366_SilicoDArT to "*Allium sativum AsFMO1* mRNA for S-allyl-L-cysteine S-oxygenase, complete cds" (GI 927028619). The same happened for *Allium* genus, where 24 sequences were found (Table 2.1). The 1,082 identified sequences were associated to 110 different species, as described in Table 2.2. Interestingly, they included eight *Allium* species, including *A. sativum*, as well as other *Liliaceae* species, such as *Asparagus officinalis*.

2.4.2. GO-term enrichment and metabolic-pathway analyses

Once genes were filtered, BLASTx analyses were performed in UniProt database. From 142 entries from BLASTx, only 120 encoded proteins (Supplementary Material 2.2). PIC values were included, maximum, minimum and mean values being 0.5, 0.005 and 0.2, respectively. These results can be slightly low for genetic-diversity studies. Notwithstanding, it should be taken into account that not all SilicoDArT markers have been taken into account for this calculation and, moreover, garlic asexual reproduction dramatically reduces the genetic variability in the species.

Regarding protein results, for those 120 hits, GO-term searches were carried out, finding 559 terms. Among them, 188, 122 and 245 belonged to the BP, CC and MF domain, respectively.

Table 2.1. Allium sativum and Allium genes showing hits after BLASTn analyses against NCBI databases. ID: DArTseq identification number; bp: base pairs; GI: gene identification (NCBI database); AN: accession number (NCBI database); Description: NCBI gene description; Species: the ones in which the gene is described; Identity: percentage of identity; and e-value: e-value score.

ID	bp	GI	AN	Description	Species	Identity (%)	e-value
9360736_SilicoDArT	69	1171486	Z69033	A. altaicum satellite DNA (strain TAX 0017, 1425, 1678, 1691)	*Allium altaicum*	91.30	1.06E-16
9334504_SilicoDArT	69	404661	L12173	Allium porrum mannose specific lectin mRNA, complete cds	*Allium ampeloprasum*	100.00	1.80E-04
9345409_SilicoDArT	59	148872676	EF633511	Allium cepa var. aggregatum lipid transfer protein 4 gene, complete cds	*Allium ascalonicum*	100.00	3.79E-21
9344172_SilicoDArT	69	1769831	Y07838	A. cepa mRNA for fructan: fructan 6G-fructosyltransferase	*Allium cepa*	95.52	1.36E-20
9344037_SilicoDArT	69	780981384	KM117265	Allium cepa 18S ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, internal transcribed spacer 2, and 26S ribosomal RNA gene, complete sequence	*Allium cepa*	97.10	2.27E-23
9344172_SilicoDArT	69	985563843	LC121826	Allium cepa 6G-FFT mRNA for fructan: fructan 6G-fructosyltransferase, complete cds, cultivar: Kitamomiji, clone: Kita1	*Allium cepa*	100.00	1.05E-26
9360914_SilicoDArT	37	469402927	AB747098	Allium cepa AcRAD21-1 mRNA for cohesin subunit RAD21-1, complete cds, cultivar: Cheonjudaego	*Allium cepa*	100.00	6.44E-09
9345409_SilicoDArT	59	171221510	EU561064	Allium cepa antimicrobial peptide mRNA, partial cds	*Allium cepa*	100.00	3.79E-21
9345409_SilicoDArT	59	2183325	AF004946	Allium cepa antimicrobial protein Ace-AMP1 precursor mRNA, complete cds	*Allium cepa*	100.00	3.79E-21
9337242_SilicoDArT	69	282767699	GU253298	Allium cepa cultivar MRSPA ATP synthase subunit 6 (atp6) gene, complete cds; and unknown gene; mitochondrial	*Allium cepa*	100.00	1.39E-05
9360464_SilicoDArT	43	741985377	KM434203	Allium cepa dihydroflavonol 4-reductase (DFR-A) gene, DFR-ADTP allele, complete cds; and transposon AcCACTA1, complete sequence	*Allium cepa*	100.00	3.85E-11
9340864_SilicoDArT	48	49781342	AY647262	Allium cepa flavonol synthase gene, complete cds	*Allium cepa*	100.00	4.94E-15
9344172_SilicoDArT	69	1081749611	KT935444	Allium cepa fructan: fructan 6G-fructosyltransferase mRNA, complete cds	*Allium cepa*	100.00	1.05E-26
9345409_SilicoDArT	59	148872670	EF633508	Allium cepa lipid transfer protein 1 gene, complete cds	*Allium cepa*	100.00	3.79E-21
9353451_SilicoDArT	69	510122042	KC466030	Allium cepa UFGT2 mRNA, complete cds	*Allium cepa*	100.00	1.05E-26
9330977_SilicoDArT	69	24460071	AB094592	Allium chinense lfs mRNA for lachrymatory factor synthase, complete cds	*Allium chinense*	94.20	4.91E-20
9345409_SilicoDArT	59	148872674	EF633510	Allium fistulosum lipid transfer protein 3 gene, complete cds	*Allium fistulosum*	100.00	3.79E-21
9325222_SilicoDArT	69	330689878	HQ738919	Allium roylei lachrymatory factor synthase (LFS) gene, partial cds	*Allium roylei*	91.30	1.06E-16
9343366_SilicoDArT	69	927028619	AB924383	Allium sativum AsFMO1 mRNA for S-allyl-L-cysteine S-oxygenase, complete cds	*Allium sativum*	92.86	2.28E-18
9341999_SilicoDArT	69	166342	M94106	Allium sativum chitinase mRNA, 3' end	*Allium sativum*	100.00	1.05E-26
9322412_SilicoDArT	69	59799342	AY766093	Allium sativum chloroplast cysteine synthase GCS2 (gcs2) mRNA, complete cds; nuclear gene for chloroplast product	*Allium sativum*	98.46	8.15E-23
9323004_SilicoDArT	69	27448223	AF384110	Allium sativum phytochelatins synthase (pcs1) mRNA, complete cds	*Allium sativum*	100.00	1.79E-09

Table 2.2. Species associated to BLASTn hits. Garlic is shown in boldface.

Species			
Agave tequilana	*Cicer arietinum*	*Lycium barbarum*	*Picea glauca*
Allium altaicum	*Cleome hassleriana*	*Lycium ruthenicum*	*Picea sitchensis*
Allium ampeloprasum	*Cucumis melo*	*Malus domestica*	*Populus euphratica*
Allium ascalonicum	*Cucumis sativus*	*Manihot esculenta*	*Populus trichocarpa*
Allium cepa	*Cyphomeris crassifolia*	*Medicago truncatula*	*Prunus mume*
Allium chinense	*Daucus carota subsp. sativus*	*Mimulus guttatus*	*Prunus persica*
Allium fistulosum	*Dimocarpus longan*	*Mirabilis jalapa*	*Prunus salicina*
Allium roylei	*Elaeis guineensis*	*Morus notabilis*	*Pyrus x bretschneideri*
Allium roylei	*Eucalyptus grandis*	*Musa acuminata subsp. malaccensis*	*Raphanus sativus*
Allium sativum	*Euphorbia esula*	*Narcissus pseudonarcissus*	*Ricinus communis*
Amborella trichopoda	*Eutrema parvulum*	*Nelumbo nucifera*	*Sesamum indicum*
Ananas bracteatus	*Eutrema salsugineum*	*Nicotiana attenuata*	*Setaria italica*
Arabidopsis lyrata subsp. lyrata	*Fragaria vesca subsp. vesca*	*Nicotiana sylvestris*	*Solanum lycopersicum*
Arabidopsis thaliana	*Fragaria x ananassa*	*Nicotiana tabacum*	*Solanum pennellii*
Arabis alpina	*Glycine max*	*Nicotiana tomentosiformis*	*Solanum tuberosum*
Arachis duranensis	*Gossypium arboreum*	*Orobanche austrohispanica*	*Sorghum bicolor*
Arachis ipaensis	*Gossypium hirsutum*	*Oryza brachyantha*	*Spirodela polyrhiza*
Asparagus officinalis	*Gossypium raimondii*	*Oryza glaberrima*	*Theobroma cacao*
Beta vulgaris subsp. vulgaris	*Hordeum vulgare subsp. vulgare*	*Oryza minuta*	*Trifolium repens*
Brachypodium sylvaticum	*Hyacinthus orientalis*	*Oryza officinalis*	*Triticum aestivum*
Brassica napus	*Ipomoea nil*	*Oryza punctata*	*Vaccinium myrtillus*
Brassica oleracea var. oleracea	*Jatropha curcas*	*Oryza rufipogon*	*Vigna angularis var. angularis*
Brassica rapa	*Juglans regia*	*Oryza sativa Indica Group*	*Vitis pseudoreticulata*
Brassica rapa subsp. chinensis	*Leea coccinea*	*Oryza sativa Japonica Group*	*Vitis vinifera*
Camelina sativa	*Linum usitatissimum*	*Panax ginseng*	*Zea mays*

Capsella rubella	*Litchi chinensis*	*Peltoboykinia tellimoides*	*Ziziphus jujuba*
Capsicum annuum	*Lotus japonicus*	*Phialophora attae*	
Carica papaya	*Lupinus angustifolius*	*Phyllostachys edulis*	

GO terms were summarized using REViGO tool, according to their domain (BP; CC, and MF; Figs. 2.1 to 2.3). Some GO terms had interesting functions. For instance, in the case of BP domain, responses to heat and oxidative stress were found (middle-left in Fig. 2.1). This could be related to entries found in Supplementary Material 2.2, such as 9360611_SilicoDArT, encoding a chaperone (Q84Q72) with response to stress function; 9345409_SilicoDArT, encoding an antimicrobial peptide (Q41258) with defense responses; 9360612_SilicoDArT, encoding a heat shock protein (P27880); and 9360611_SilicoDArT (P19037). Finally, another sequence with interesting functions was 9357227_SilicoDArT, encoding a protein with responses to oxidative stress and heme-group binding (B9R8E4). BP related to transcription and translation were also found (bottom of Fig. 2.1). Additionally, metabolic processes, such as lipid or photosynthesis metabolism, were found in the central part of the graph. Finally, in the bottom-left corner, processes related to hormone or nutrient response were shown.

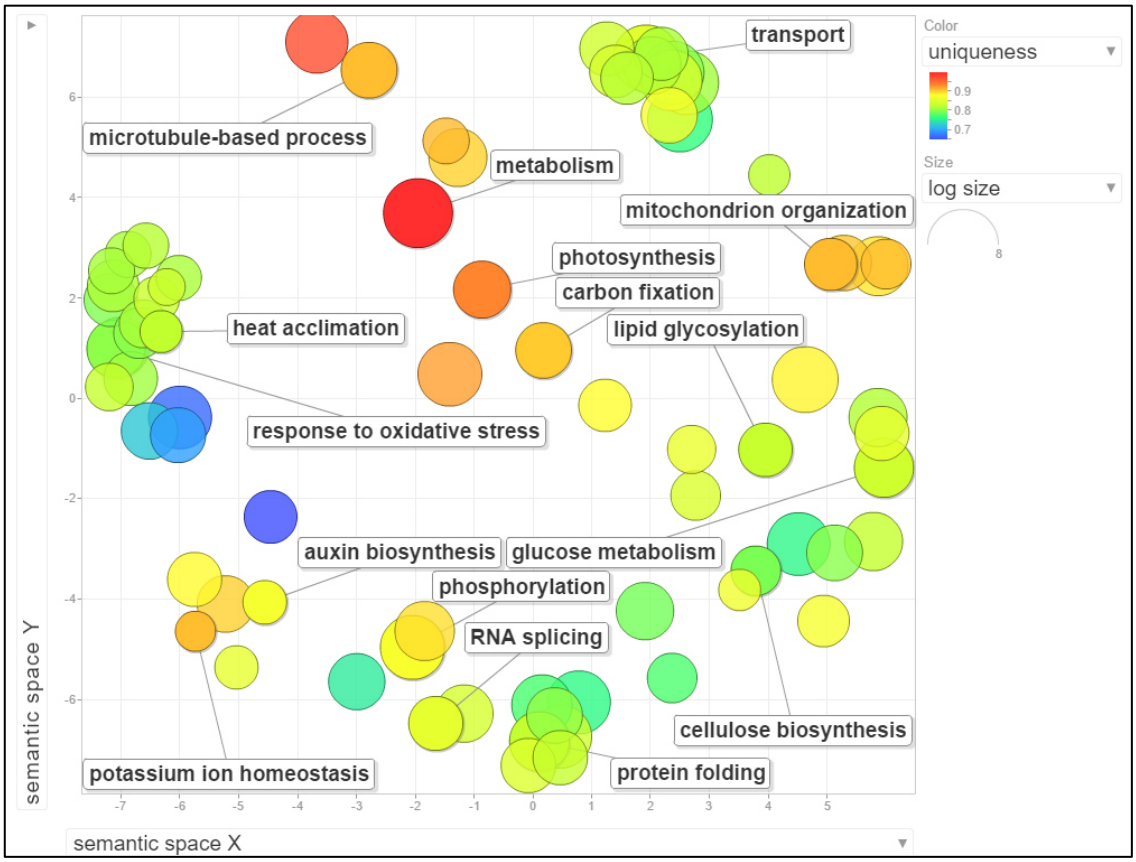


Figure 2.1. REViGO graph summarizing GO terms, according to biological processes. Main functions are shown. Dot color codes correspond to degree of uniqueness, from 0 (not unique at all) to 1 (totally unique). Dot dimensions are log sizes; that is, the logarithmic number of genes annotated within the terms.

In relation to CC domain, locations related to BP and MF domains were included. It is worth mentioning that both, uniqueness and log size, showed that most processes belonged to nucleus, ribosome or chloroplastic regions.

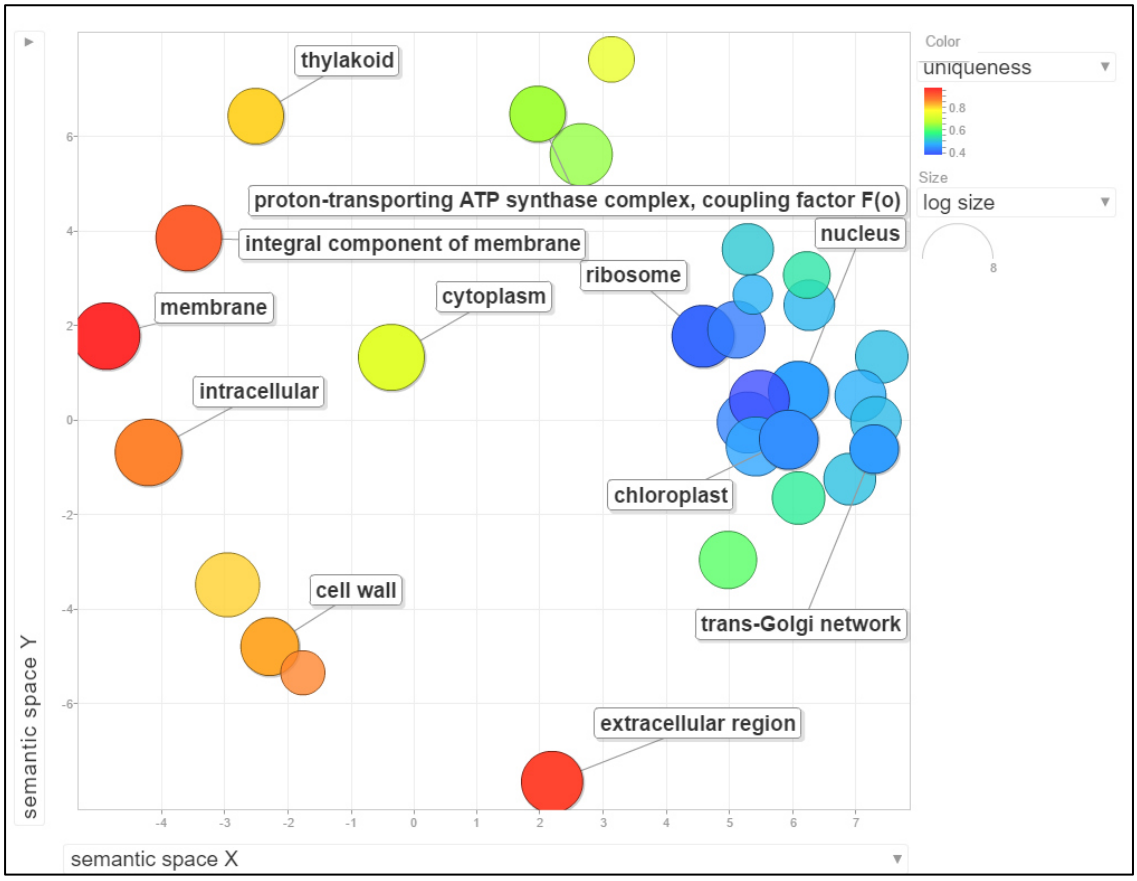


Figure 2.2. REViGO graph summarizing GO terms, according to cellular components. Main functions are included. See legend of Fig. 2.1.

Additionally, MF domain (Fig. 2.3) showed mainly enzyme functions. Dots could be divided into four main groups. On top, those related to nucleotide or protein binding are shown. On each side are displayed, enzymatic processes for protein modification. Protein binding and lipid and transcription-cofactor activity are exhibited on middle part. The most unique functions are located in the middle.

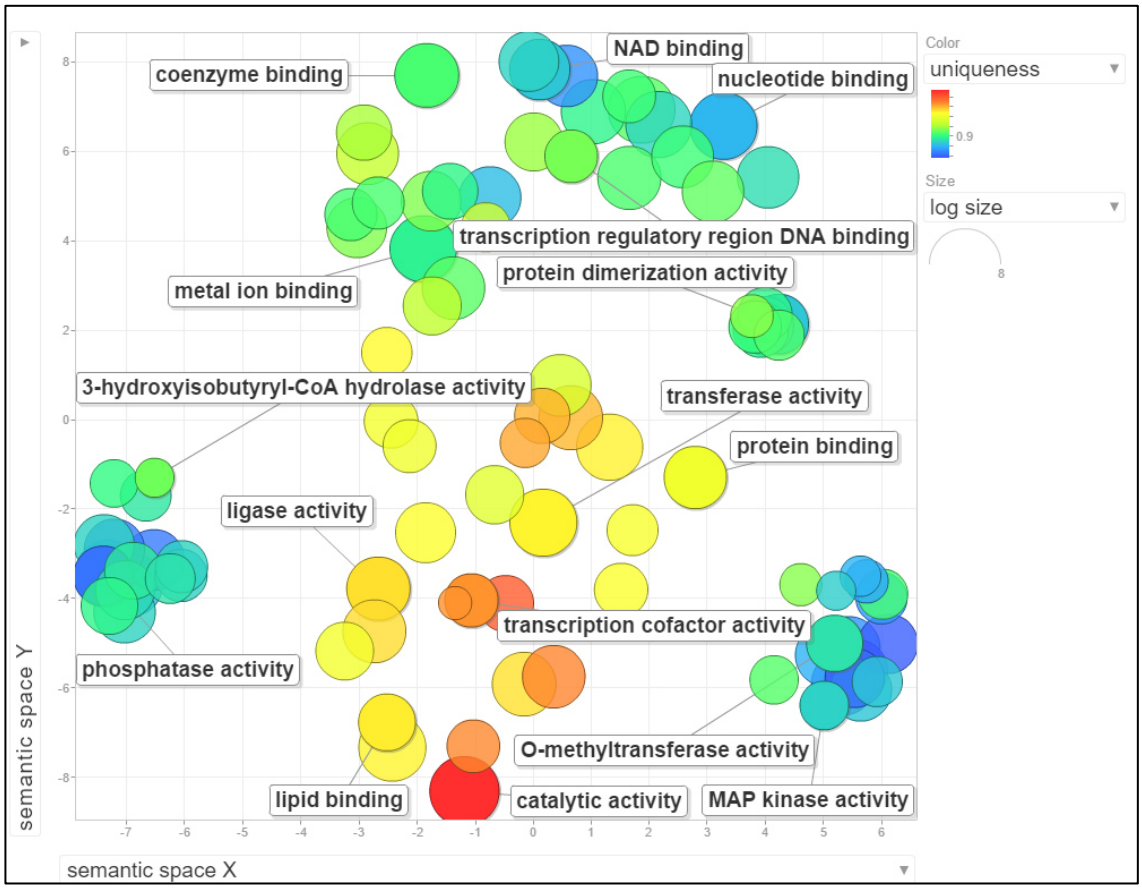


Figure 2.3. REViGO graph summarizing GO terms, according to molecular function. Main functions are displayed. See legend of Fig. 2.1.

In addition to GO analyses, GO slim and metabolic pathway evaluation was done in PANTHER. First, all UniProt codes were changed to *A. thaliana* ones looking for the most similar homologous sequence. In all cases, chosen proteins were manually reviewed in SWISS-PROT database (Bairoch and Apweiler, 2000), except for two entries (A0A178UQ69 and A0A178V6V7) that did not have any reviewed accession. Conversely, not all proteins had an *A. thaliana* correspondence. Hence, from the 120 proteins found in UniProt database for different species, only 94 had an *A. thaliana* homologous protein (Supplementary Material 2.2). Several graphs were taken from PANTHER website, in order to illustrate these results (Figs. 2.4 to 2.7).

Fig. 2.4. shows the main types of proteins found in PANTHER database. In total, from the 94 *A. thaliana* codes, PANTHER found 87 correspondences. Among them, there were 58 protein classes. Main classes were: calcium-binding protein (PC00060), cell-junction protein (PC00070), chaperone (PC00072), cytoskeletal protein (PC00085), enzyme modulator (PC00095), hydrolase (PC00121), isomerase (PC00135), ligase (PC00142), lyase (PC00144), nucleic-acid binding (PC00171), oxidoreductase (PC00176), transcription factor (PC00218), transferase (PC00220), and transporter (PC00227).

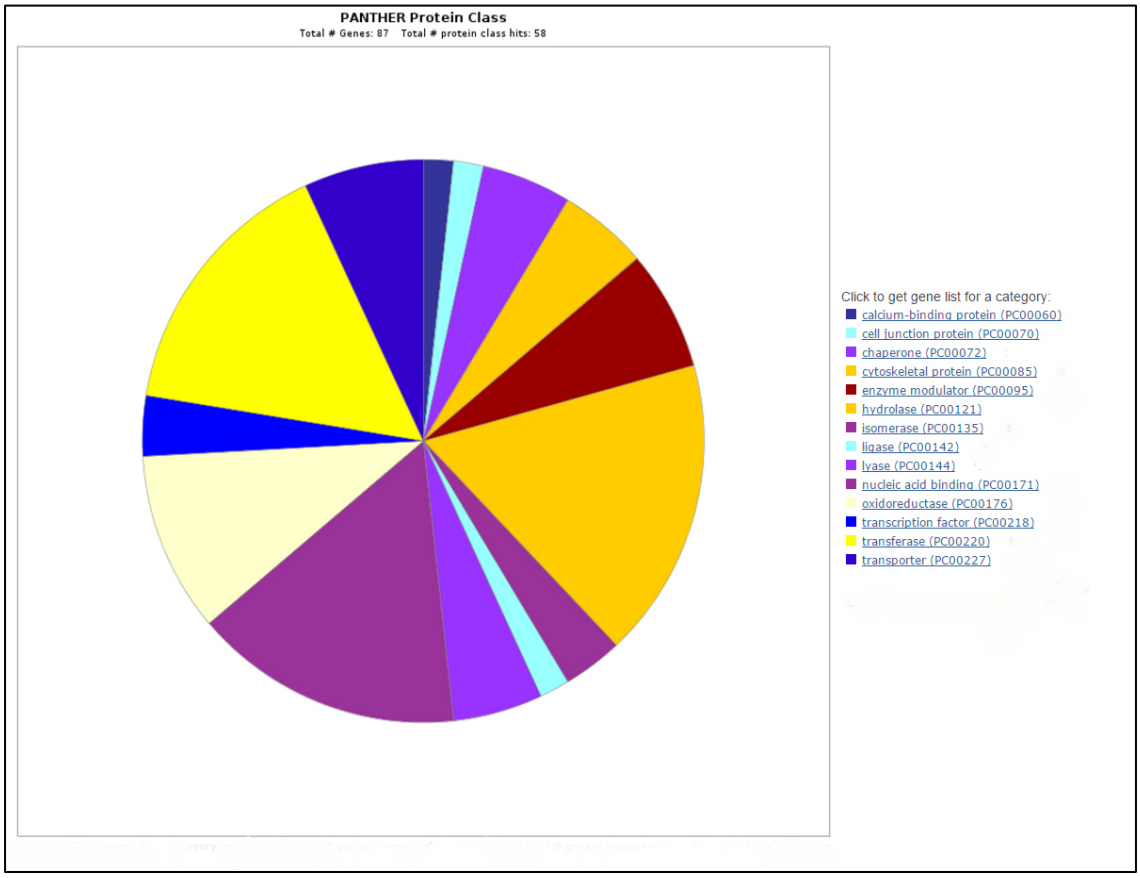


Figure 2.4. 14 main protein classes found in PANTHER database. A total of 87 genes were analyzed, finding 58 protein-class hits. Each main class has a unique color legend.

Regarding GO-slim data, one graph was generated for each GO domain. In total, 232 terms were found for the three domains. GO slim for BP showed 104 hits for the 87 proteins found in PANTHER database (Fig. 2.5). Eight main BP terms are shown: biological regulation (GO:0065007), cellular component organization or biogenesis (GO:0071840), cellular process (GO:0009987), developmental process (GO:0032502), localization (GO:0051179), metabolic process (GO:0008152), multicellular organismal process (GO:0032501), and response to stimulus (GO:0050896).

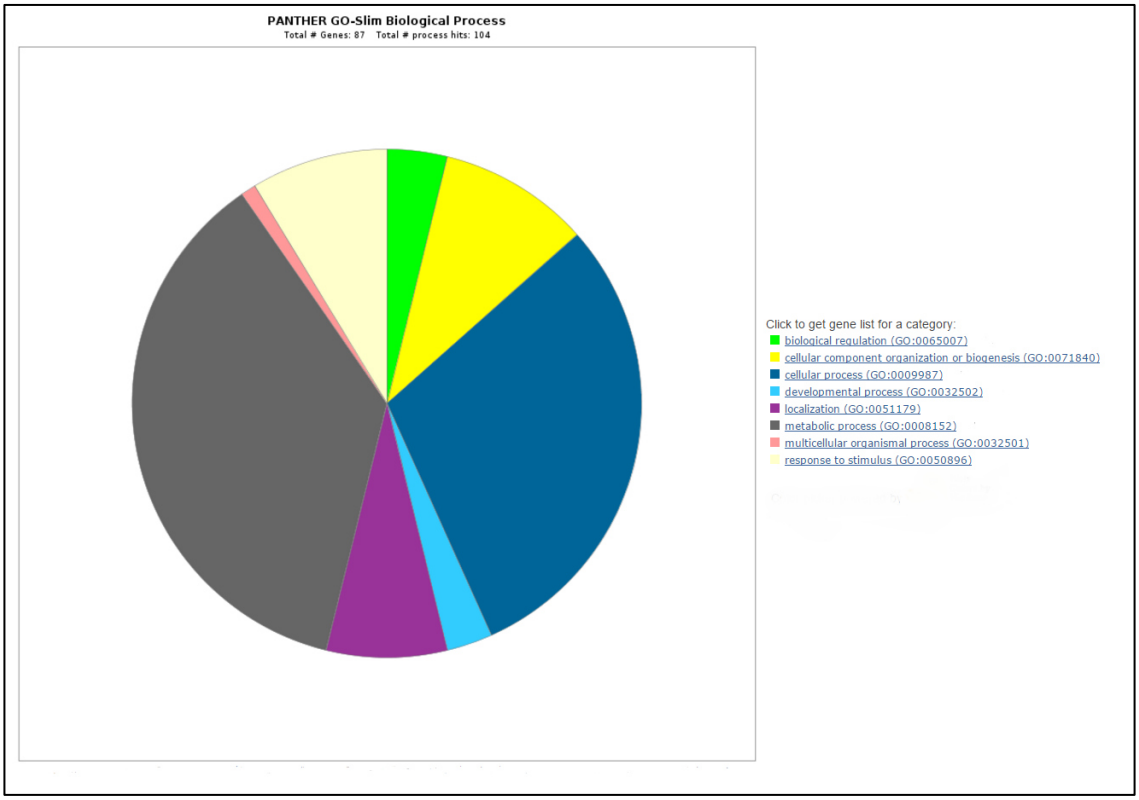


Figure 2.5. Eight main GO-slim terms for biological process domain in PANTHER database. A total of 87 genes were analyzed, finding 104 hits. Each main category has a unique color legend.

GO slim for CC found 63 hits (Fig. 2.6). In this case, six main categories are shown: cell junction (GO:0030054), cell part (GO:0044464), extracellular region (GO:0005576), macromolecular complex (GO:0032991), membrane (GO:0016020), and organelle (GO:0043226).

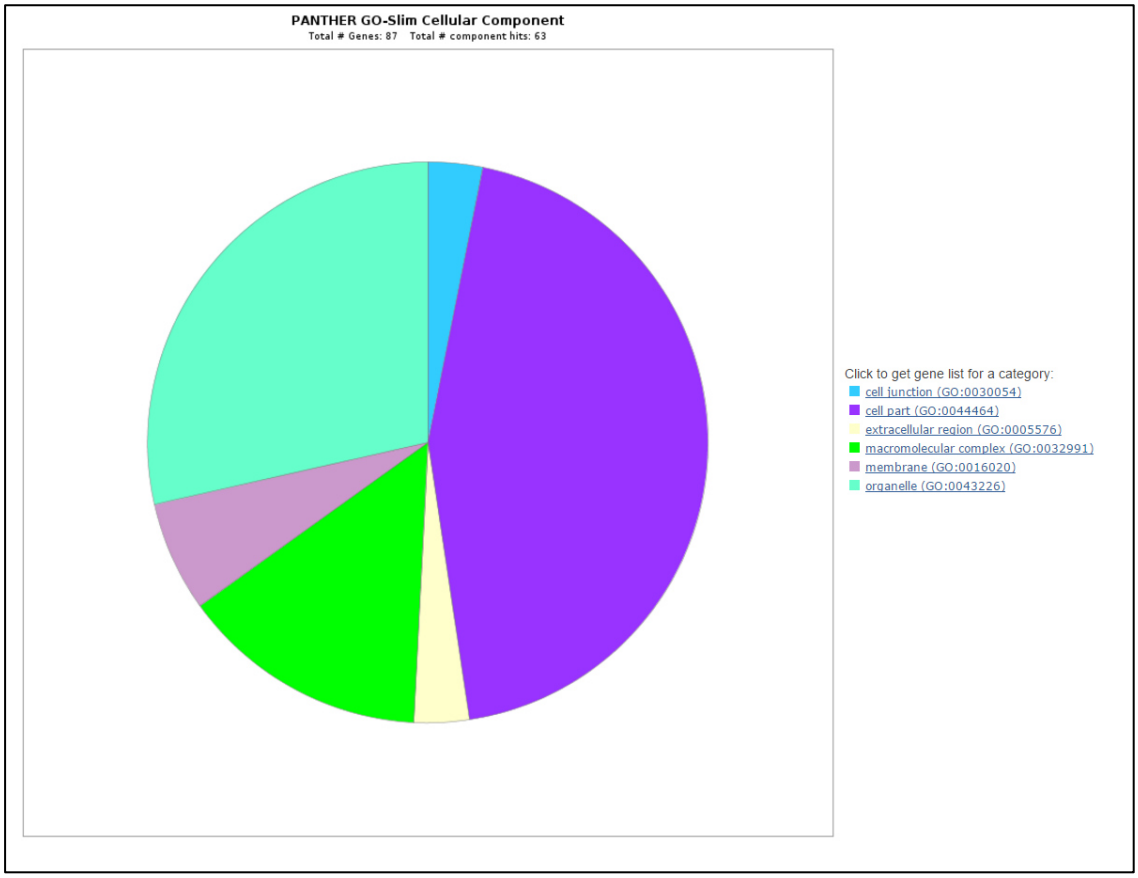


Figure 2.6. Eight main GO-slim terms for cellular component domain in PANTHER database. A total of 87 genes were analyzed, finding 63 hits. Each main category has a unique color legend.

GO slim for MF found 65 hits (Fig. 2.7). Six main categories are shown: antioxidant activity (GO:0016209), binding (GO:0005488), catalytic activity (GO:0003824), structural molecule activity (GO:0005198), translation regulator activity (GO:0045182), and transporter activity (GO:0005215).

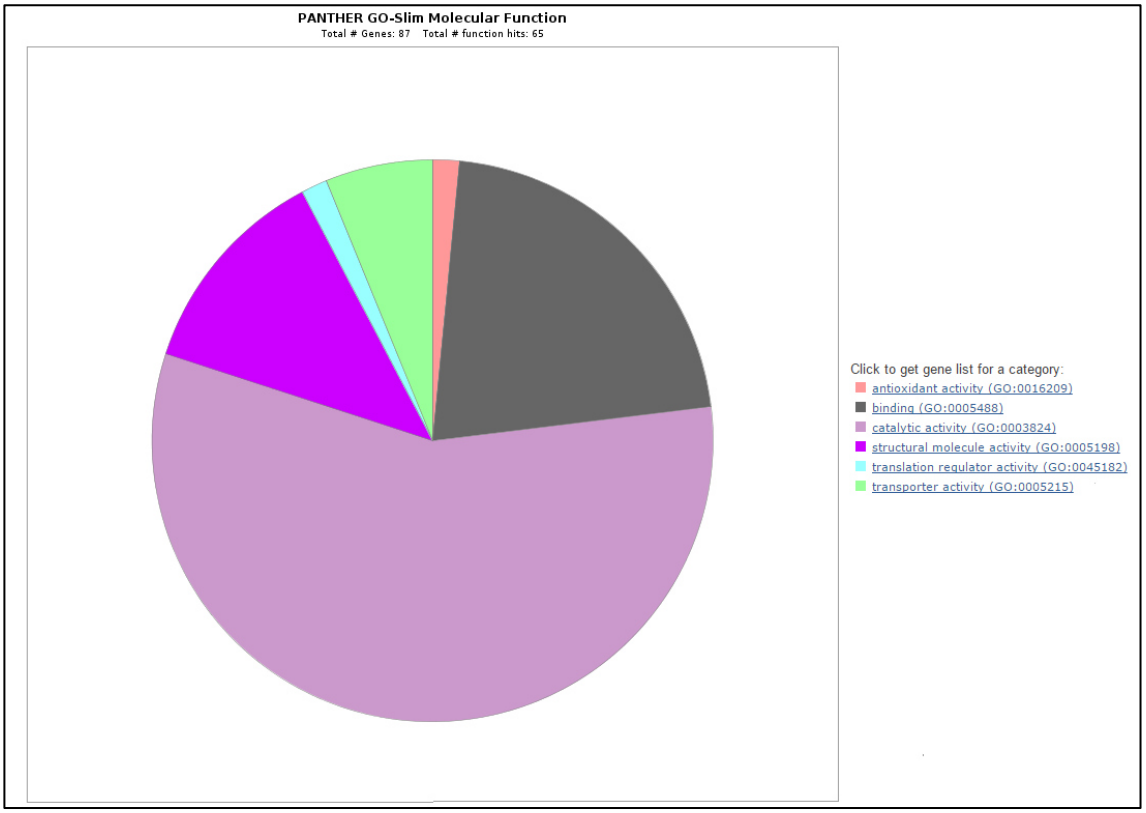


Figure 2.7. Eight main GO-slim terms for molecular function domain in PANTHER database. A total of 87 genes were analyzed, finding 65 hits. Each main category has a unique color legend.

In relation to metabolic pathways, the following results were found. From the 82 *A. thaliana* genes found in PANTHER, 11 metabolic pathways were revealed. Ten main categories are shown in the histogram: ATP synthesis (P02721), ascorbate degradation (P02729), cholesterol biosynthesis (P00014), coenzyme A biosynthesis (P02736), fibroblast Growth Factor (FGF) signaling pathway (P00021), fructose galactose metabolism (P02744), glycolysis (P00024), nicotinic acetylcholine receptor signaling pathway (P00044), TriCarboxylic Acid (TCA) cycle (P00051), and ubiquitin proteasome pathway (P00060). Only one gene was involved in each pathway, except for the case of the fructose galactose metabolism (P02744), which had two.

Finally, a graph including all 188 BP GO terms was generated with Blast2GO (Supplementary Material 2.3). Due to the large size of these graphs, only BP GO terms are shown. BP domain was chosen since, after GO-slim analyses, it was the one with less redundancy among its terms. In this graph, all biological processes were related among them by "is a", "part of", "regulates" or "positively regulates". Processes related to response to stimulus are shown on left part of graph. Among them, there are internal or external stimuli, and also responses to stress processes. In total, 25 GO terms were associated to response to stimulus (GO:0050896), and five belonged specifically to response to stress (GO:0006950). Immediately after, cellular processes are described, with some regulations between them. 10 main GO terms for cellular processes (GO:0009987) were subdivided into 47 processes. Regulations in this category were found; for instance, positive regulation of cellular metabolic process (GO:0031325) positively regulated Cellular metabolic process (GO:0044237). Likewise, positive regulation of cellular process (GO:004852); and Regulation of cellular process (GO:0050794) regulated cellular processes (GO: :0009987). Regulation of cellular metabolic process (GO:0031323) was involved in this regulation as well.

In the central part of the graph there are other signaling, developmental, single or multicellular, as well as metabolic processes. A total of 10 main processes were involved in metabolic process (GO:0008152), subdivided into a total of 39 subprocesses. Five regulations were found here. Positive regulation of biosynthetic process (GO:0009891) positively regulated biosynthetic process (GO:0009058). On the other hand, organic

substance metabolic process (GO:007704) was regulated by regulation of biosynthetic process (GO:0009889) and positive regulation of nitrogen compound metabolic process (GO:0051173). Additionally, primary metabolic process (GO:0044238) was regulated by regulation of nitrogen compound metabolic process (GO:0051173), and regulation of primary metabolic process (GO:0080090). Biological regulation processes were located on the right part, having four main terms and 10 subprocesses. On the other hand, there were two kinds of processes on the right end: those related to cellular localization, with four main terms and 11 subterms, and one relate to cell killing (GO:0001906), with only one term, as killing of cells of other organisms (GO:0031640).

2.5. Discussion

The applicability of virtual genomic tools to assess non-model species at a genetic level has a great relevance. This is mostly due to advances in sequencing technologies, bioinformatics processing, and open databases. Due to the evolutionary process, unknown genomic/proteomic regions of a particular species may be conserved and, therefore, show identities to other described species in the databases. This way, finding identities and putative functions in unknown sequences is feasible, even for species without reference genome (Bellin et al., 2009) like garlic. Hence, by comparing sequences using online alignment tools such as BLAST on databases like the ones at NCBI or UniProt, sequences of species without prior information can be identified (Horan et al., 2008). This strategy broadens the understanding of genetic diversity across different species (Xianjun et al., 2014). Indeed, finding correspondence to described *Allium* genes and new ones is crucial to broaden our knowledge about garlic genetics. In fact, unknown sequences may become more relevant than known ones, once they are properly identified (Horan et al., 2008). However, caution should be taken when using these approaches. Firstly, errors increase with evolutionary distance. Secondly, results may be biased towards conserved genes (Hornett and Wheat, 2012).

After DArTseq analyses in a previous work, 33,423 short sequences of garlic genome were generated. In this one, BLASTn and BLASTx analyses have been performed in order to further identify genes and proteins present in these sequences, in the absence of reference genome. In addition, annotation of GO terms has been carried out. Fortunately, DArTseq is a technique that allows the study of genetic diversity in non-model species, with scarce or absent genetic information. By using an *in-silico* approach, 1,082 genes from 110 different species were detected. After filtering for duplicated accessions, identity and “e-value” scores, 142 genes remained, encoding 120 proteins. GO annotations resulted in 559 terms. Finally, 82 *A. thaliana* genes were found in PANTHER, belonging to eleven metabolic pathways. The obtained results are interesting and can be used as a starting point for further genetic or sequencing studies (Kim et al., 2009; da Cunha et al., 2014).

This way, BLASTn analyses revealed some interesting genes (Supplementary Material 2.1). The first case is 9354508_SilicoDArT sequence. One hit with *Sesamum indicum* L. was found for this marker. Specifically, it had identity to GI:747066802, whose description is "PREDICTED: *Sesamum indicum* cellulose synthase A catalytic subunit 4 [UDP-forming] (LOC105163451), mRNA". This may have industrial interest, as previous works have already reported for cellulases (Kim et al., 2010), being both related to cellulose metabolism. Another sequence described in chapter 1 that has been found here was 9334504_SilicoDArT (GI:404661), corresponding to "*Allium porrum* mannose specific lectin mRNA, complete cds" (Smeets et al., 1997). Such information is useful to identify polymorphisms and to develop molecular markers in sequences or genes of interest, as described (Dorado, Besnard, et al., 2015; Dorado, Unver, et al., 2015).

Regarding *Allium* metabolism specifically, 18 different coding genes were found (Table 2.1). Six different synthases were identified: i) *Allium sativum* chloroplast cysteine synthase GCS2 (*gcs2*) mRNA, complete CDS; nuclear gene for chloroplast product (9322412_SilicoDArT, GI:59799342); ii) *Allium sativum* phytochelatins synthase (*pcs1*) mRNA, complete CDS (9323004_SilicoDArT, GI:27448223); iii) *Allium roylei* lachrymatory factor synthase (*LFS*) gene, partial CDS (9325222_SilicoDArT, GI:330689878); iv) *Allium chinense lfs* mRNA for lachrymatory factor synthase, complete CDS (9330977_SilicoDArT, GI:24460071); v) *Allium cepa* cultivar MRSPA ATP synthase subunit 6 (*atp6*) gene, complete CDS, and unknown gene, mitochondrial (9337242_SilicoDArT, GI:282767699); and vi) *Allium cepa* flavonol synthase gene, complete CDS (9340864_SilicoDArT, GI:49781342). This information could be interesting to seek polymorphism in this compounds in garlic varieties.

Other *Allium* enzymes were: i) *Allium sativum* chitinase mRNA, 3' end (9341999_SilicoDArT, GI:166342); ii) *Allium sativum AsFMO1* mRNA for S-allyl-L-cysteine S-oxygenase, complete CDS (9343366_SilicoDArT, GI:927028619); iii), *A. cepa* mRNA for fructan:fructan 6G-fructosyltransferase (9344172_SilicoDArT, GI:1769831); iv) *Allium cepa* UFGT2 (glycosyltransferase) mRNA, complete CDS (9353451_SilicoDArT, GI:510122042); v) *Allium cepa* var. *aggregatum* lipid transfer protein 4 gene, complete CDS (9345409_SilicoDArT, GI:148872676); and vi) *Allium*

cepa antimicrobial protein *Ace-AMP1* precursor mRNA, complete CDS for this same SilicoDArT entry (GI 2183325); vii) *Allium cepa* dihydroflavonol 4-reductase (*DFR-A*) gene, *DFR-ADTP* allele, complete CDS (9360464_SilicoDArT, GI:741985377); viii) transposon AcCACTA1, complete sequence (9334504_SilicoDArT, GI:404661); ix) *Allium porrum* mannose specific lectin mRNA, complete CDS (9344037_SilicoDArT, GI:780981384); x) *Allium cepa* 18S ribosomal RNA gene, internal transcribed spacer 1; 5.8S ribosomal RNA gene, internal transcribed spacer 2; and 26S ribosomal RNA gene, complete sequence; and xi), *Allium cepa AcRAD21-1* mRNA for cohesin subunit RAD21-1, complete CDS, cultivar: Cheonjudaego (9360914_SilicoDArT, GI:469402927).

On the other hand, it is vital to have information as broad as possible about biological mechanisms to adapt and tolerate biotic stresses, as well as abiotic ones, in the current trend of climate change and global warming. Identifying involved genes and their polymorphisms, as described (Dorado, Besnard, et al., 2015; Dorado, Unver, et al., 2015), can help to find strategies in order to study responses to such stresses in garlic. An interesting set was a group of genes with response to stress functions, that appeared in the middle-left area of Fig. 2.1 and in Supplementary material 2.1. They include several hits: i) *Agave tequilana* F.A.C. Weber gene, encoding chaperone (Q84Q72), with response to stress (9360611_SilicoDArT, GI:99033682); ii) *Allium cepa* L. gene encoding antimicrobial peptide (Q41258) with defense responses (9345409_SilicoDArT, GI:171221510); iii) *Ipomoea nil* (L.) Roth gene encoding another heat-shock (*P27880*) protein (9360612_SilicoDArT, GI:1109265466); iv), *Nicotiana tabacum* L. gene, also encoding a heat-shock (P19037) protein (9360611_SilicoDArT, GI:662247390); and v) *Ricinus communis* L. gene encoding protein (B9R8E4) with responses to oxidative stress and heme-group binding (9357227_SilicoDArT, GI:1000986181).

Another interesting group of genes involved transcription factors, which are known to be key regulators of gene expression (Supplementary material 2.1): i) *Daucus carota* subsp. s*ativus* (Hoffm.) Schübl. & G. Martens gene, described as “PREDICTED: *Daucus carota* subsp. *sativus* ethylene-responsive transcription factor ERF017-like (*LOC108213718*), mRNA” (9325878_SilicoDArT, GI:1040860766); ii) *Gossypium arboreum* L. gene, considered as “PREDICTED: *Gossypium arboreum* transcription

factor JUNGBRUNNEN 1-like (*LOC108463429*), mRNA" (9334524_SilicoDArT, GI:1050594794); iii) *Nicotiana tabacum* L. gene, recorded as "PREDICTED: *Nicotiana tabacum* transcription factor MYB26-like (*LOC107819474*), transcript variant X2, mRNA" (9351778_SilicoDArT, GI:1025307851); iv) *Phoenix dactylifera* L. genome, as "PREDICTED: *Phoenix dactylifera* AP2-like ethylene-responsive transcription factor AIL5 (*LOC103717110*), mRNA" (9339167_SilicoDArT, GI:1052192815), with interesting response to such hormone; and v) *Populus euphratica* Oliv. gene, which is described as "PREDICTED: *Populus euphratica* transcription factor bHLH68 (*LOC105142436*), transcript variant X2, mRNA" (9328493_SilicoDArT, GI:743909762).

In addition, accessions related to DNA transcription or translation were found: i) *Asparagus officinalis MSH1* mRNA, partial CDS, corresponding to gene encoding protein for DNA-mismatch repair (9352361_SilicoDArT, GI:700253107); ii) *Gossypium raimondii*, identified as "PREDICTED: *Gossypium raimondii* DNA mismatch repair protein MSH1, mitochondrial (*LOC105800651*), mRNA" (9331040_SilicoDArT, GI:823179024); iii) *Cucumis melo* related to chromatin remodeling, described as "PREDICTED: *Cucumis melo* protein CHROMATIN REMODELING 4 (*LOC103484261*), transcript variant X4, mRNA" (9347750_SilicoDArT, GI:1035395921); and iv) *Juglans regia* RNA polymerase, considered "PREDICTED: *Juglans regia* RNA polymerase II-associated protein 3 (*LOC108987762*), transcript variant X3, mRNA" (9354847_SilicoDArT, GI:1098820251).

Additionally, two groups of protein- and carbohydrate-modifying enzymes were also found (Supplementary Material 2.1). For proteins, genes were: i) *Arabidopsis lyrata* subsp. *lyrata* protein translocase subunit secA chloroplast precursor, mRNA (9353780_SilicoDArT, GI:297809984); ii) *Beta vulgaris* subsp. *vulgaris*, shown as "PREDICTED: *Beta vulgaris* subsp. *vulgaris* RNA-binding protein 25 (*LOC104897399*), transcript variant X3, mRNA" (9333976_SilicoDArT, GI:1108943899); iii) *Carica papaya*, recorded as "*Carica papaya* GTP-binding nuclear protein (RAN) mRNA, partial CDS" (9331630_SilicoDArT, GI:1121551646); iv) *Cicer arietinum*, identified as "PREDICTED: *Cicer arietinum* glycogen synthase kinase-3 homolog MsK-1

(*LOC101500512*), transcript variant X3, mRNA" (9341417_SilicoDArT, GI:828303965); v) *Juglans regia*, included as "PREDICTED: *Juglans regia* serine/threonine-protein phosphatase BSL3-like (*LOC109018379*), mRNA" (9336922_SilicoDArT, GI:1098789479); vi) *Juglans regia*, found as "PREDICTED: *Juglans regia* casein kinase 1-like protein 1 (*LOC109009397*), mRNA" (9354458_SilicoDArT, GI:1098721498); vii) *Prunus mume*, stored as "PREDICTED: *Prunus mume* E3 ubiquitin-protein ligase RHF2A (*LOC103329838*), mRNA" (9352365_SilicoDArT, GI:1027091046); and viii) *Vitis vinifera*, considered as "PREDICTED: *Vitis vinifera* mitogen-activated protein kinase YODA (*LOC100257467*), transcript variant X3, mRNA" (9335110_SilicoDArT, GI:1104680471).

Hits for carbohydrates were as follows: i) *Cucumis sativus* probable sucrose-phosphate synthase 2 (*LOC101208942*), mRNA (9349862_SilicoDArT, GI:793420965); ii) *Cyphomeris crassifolia* voucher Douglas 2203 clone 1 phosphoenolpyruvate carboxylase (*ppc-1E1*) gene, partial CDS (9325282_SilicoDArT, GI:664680186); iii) PREDICTED: *Elaeis guineensis* glucose-6-phosphate 1-dehydrogenase, chloroplastic-like (*LOC105051009*), transcript variant X2, mRNA (9341988_SilicoDArT, GI:1130672164); iv) PREDICTED: *Eucalyptus grandis* fructose-bisphosphate aldolase, cytoplasmic isozyme 1 (*LOC104449094*), mRNA (9330263_SilicoDArT, GI:1091481656); v) PREDICTED: *Musa acuminata* subsp. malaccensis serine/threonine-protein phosphatase BSL3-like (*LOC103982958*), transcript variant X6, mRNA (9330984_SilicoDArT, GI:1091693398); vi) PREDICTED: *Sesamum indicum* malate dehydrogenase (*LOC105171782*), mRNA (9342092_SilicoDArT, GI:747087535); vii) PREDICTED: *Solanum pennellii* probable beta-1,3-galactosyltransferase 2 (*LOC107023985*), transcript variant X2, mRNA (9338024_SilicoDArT, GI:970037987); and viii) PREDICTED: *Vitis vinifera* beta-galactosidase 3 (*LOC100232848*), mRNA (9355596_SilicoDArT, GI:1105486800), among others.

In relation to GO and GO-slim terms, it is interesting to highlight that, on average, four terms were found for each protein hit. Most of them were for MF domain (245), followed by BP (188), and finally CC having 122. On the other hand, in the case of GO

slim, where 232 hits were detected, the figures were different. Thus, BP was the domain with the largest number of hits (104), followed by MF (65), with only a difference of two in comparison to CC (63). This could suggest that the diversity of biological processes found in GO was bigger than for molecular functions. In general, these processes and functions were related to DNA modification, DNA and RNA transcription and translation, protein and carbohydrate metabolism, and responses to stresses. One last remarkable fact regarding metabolic pathways is that, for each of them, only one gene was involved, except in the case of Fructose galactose metabolism (P02744), in which two were found.

Finally, some aspects must be remarked as well. Initially, 142 genes were found in NCBI database. Conversely, only 120 had a protein correspondence in UniProt database after BLASTx analyses. This is probably due to the fact that some genes are transcribed but not translated, or simply due to the lower representation of peptide databases in relation to nucleic acids. As expected, information was also missing when UniProt codes were switched exclusively to *A. thaliana* ones. For instance, *Allium lacrimatory* factors that are specific of this species, did not have a correspondence in the former. Indeed, these issues are likely to happen when non-model species are used (Horan et al., 2008). In any case, this is the only approach to find genetic or metabolic information for a species like garlic, without reference genome.

2.6. Conclusions

A total of 33,423 short-polymorphic sequences resulting from DArTseq analyses have been used, in order to identify genes present in garlic, in the absence of reference genome. BLAST, GO, and metabolic pathways analyses have helped to shed light in the large and expected complex genome of this asexually-reproducing species. Although some information was missing, due to the lack of identity with available information in available databases, data found here could be of useful for further genetic and genomic studies in garlic. This includes the development of polymorphic molecular markers in sequences or genes of interest. Likewise, the identification of enzymes with industrial interest like cellulase. Finally, other relevant genes were those involved in adaptation and defense against both biotic and abiotic stresses.

2.7. References

ADAMS, M.D., S.E. CELNIKER, R.A. HOLT, ET AL. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185–2195.

ALTSCHUL, S.F., W. GISH, W. MILLER, E.W. MYERS, and D.J. LIPMAN. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

BAIROCH, A., and R. APWEILER. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* 28: 45–48.

BELLIN, D., A. FERRARINI, A. CHIMENTO, O. KAISER, N. LEVENKOVA, P. BOUFFARD, and M. DELLEDONNE. 2009. Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model species. *BMC Genomics* 10: 555–563.

CONESA, A., S. GÖTZ, J.M. GARCÍA-GÓMEZ, J. TEROL, M. TALÓN, and M. ROBLES. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674–3676.

CRUZ, V.M.V., A. KILIAN, and D.A. DIERIG. 2013. Development of DArT marker platforms and genetic diversity assessment of the us collection of the new oilseed crop Lesquerella and related species. *PloS One* 8: 5.

DA CUNHA, C.P., F.V. RESENDE, M.I. ZUCCHI, and J.B. PINHEIRO. 2014. SSR-based genetic diversity and structure of garlic accessions from Brazil. *Genetica* 142: 419–431.

DORADO, G., G. BESNARD, T. UNVER, and P. HERNÁNDEZ. 2015. Polymerase Chain Reaction (PCR), *In* Caplan M [ed]. Reference Module in Biomedical Sciences. Biochemistry, Cell Biology and Molecular Biology. Elsevier Amsterdam.

DORADO, G., S. GÁLVEZ, H. BUDAK, T. UNVER, and P. HERNÁNDEZ. 2015. Nucleic-acid sequencing. *In* Caplan M [ed]. Reference Module in Biomedical Sciences. Biochemistry, Cell Biology and Molecular Biology. Elsevier Amsterdam.

DORADO, G., T. UNVER, H. BUDAK, and P. HERNÁNDEZ. 2015. Molecular markers. *In* Caplan M [ed]. Reference Module in Biomedical Sciences. Biochemistry, Cell Biology and Molecular Biology. Elsevier Amsterdam.

GARAVITO, A., C. MONTAGNON, R. GUYOT, and B. BERTRAND. 2016. Identification by the DArTseq method of the genetic origin of the Coffea canephora cultivated in Vietnam and Mexico. *BMC Plant Biology* 16: 242.

HARRIS, M.A., J. CLARK, A. IRELAND, ET AL. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 32: D258-261.

HORAN, K., C. JANG, J. BAILEY-SERRES, R. MITTLER, C. SHELTON, J.F. HARPER, J.-K. ZHU, J. C. CUSHMAN, M. GOLLERY, and T. GIRKE. 2008. Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiology* 147: 41–57.

HORNETT, E.A., and C.W. WHEAT. 2012. Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC Genomics* 13: 361.

KIM, A., R. KIM, D. KIM, S. CHOI, A. KANG, S. NAM, and H. PARK. 2010. Identification of a novel garlic cellulase gene. *In* Plant Molecular Biology Reporter, 388–93.

KIM, D.-W., T.-S. JUNG, S.-H. NAM, H.-R. KWON, A. KIM, S.-H. CHAE, S.-H. CHOI, D-W. KIM, R. N. KIM, and H-S. PARK. 2009. GarlicESTdb: an online database and mining tool for garlic EST sequences. *BMC Plant Biology* 9: 61.

MI, H., A. MURUGANUJAN, J.T. CASAGRANDE, and P.D. THOMAS. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols* 8: 1551–1566.

SMEETS, K., E.J.M.V. DAMME, P. VERHAERT, A. BARRE, P. ROUGÉ, F.V. LEUVEN, and W.J. PEUMANS. 1997. Isolation, characterization and molecular cloning of the mannose-binding lectins from leaves and roots of garlic (*Allium sativum* L.). *Plant Molecular Biology* 33: 223–234.

SUPEK, F., M. BOŠNJAK, N. ŠKUNCA, and T. ŠMUC. 2011. REVIGO Summarizes and visualizes long lists of gene ontology terms. *PloS One* 6: e21800.

THE UNIPROT CONSORTIUM. 2016. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45: D158–D169.

XIANJUN, P., T. LINHONG, W. XIAOMAN, W. YUCHENG, and S. SHIHUA. 2014. *De novo* assembly of expressed transcripts and global transcriptomic analysis from seedlings of the paper mulberry (*Broussonetia kazinoki* x *Broussonetia papyifera*). *PloS One* 9: e97487.

2.8. Supplementary Material

Supplementary Material 2.1. See legend of Table 2.1.

ID	bp	GI	AN	Description	Species	Identity (%)	e-value
9321425_SilicoDArT	69	1109138320	XM_010517534	PREDICTED: Camelina sativa 29 kDa ribonucleoprotein, chloroplastic (LOC104791610), mRNA	*Camelina sativa*	85.71	2.32E-08
9321555_SilicoDArT	69	1026026375	XM_016703973	PREDICTED: Capsicum annuum 40S ribosomal protein S4-like (LOC107859079), mRNA	*Capsicum annuum*	92.00	1.79E-09
9322412_SilicoDArT	69	59799342	AY766093	Allium sativum chloroplast cysteine synthase GCS2 (gcs2) mRNA, complete cds; nuclear gene for chloroplast product	*Allium sativum*	98.46	8.15E-23
9322838_SilicoDArT	69	828319638	XM_004505058	PREDICTED: Cicer arietinum rac-like GTP-binding protein RAC13 (LOC101498092), mRNA	*Cicer arietinum*	91.11	1.08E-06
9323004_SilicoDArT	69	27448223	AF384110	Allium sativum phytochelatins synthase (pcs1) mRNA, complete cds	*Allium sativum*	100.00	1.79E-09
9323004_SilicoDArT	69	27448223	AF384110	Allium sativum phytochelatins synthase (pcs1) mRNA, complete cds	*Allium sativum*	100.00	1.79E-09
9323284_SilicoDArT	69	1117415913	XM_019558619	PREDICTED: Lupinus angustifolius auxin transporter-like protein 5 (LOC109325971), transcript variant X3, mRNA	*Lupinus angustifolius*	88.06	2.97E-12
9323367_SilicoDArT	69	1102790753	XM_010245323	PREDICTED: Nelumbo nucifera myosin-6-like (LOC104587632), mRNA	*Nelumbo nucifera*	94.29	1.80E-04
9323600_SilicoDArT	54	662247392	KJ719266	Nicotiana tabacum HSP19.8 mRNA, partial cds	*Nicotiana tabacum*	97.06	1.39E-05
9324289_SilicoDArT	53	802555686	XM_012210184	PREDICTED: Jatropha curcas polygalacturonase-like (LOC105628710), mRNA	*Jatropha curcas*	100.00	3.87E-06
9324822_SilicoDArT	37	733578762	LN713258	Cucumis melo genomic chromosome, chr_4	*Cucumis melo*	100.00	6.48E-04
9325222_SilicoDArT	69	330689878	HQ738919	Allium roylei lachrymatory factor synthase (LFS) gene, partial cds	*Allium roylei*	91.30	1.06E-16
9325282_SilicoDArT	69	664680186	KJ161694	Cyphomeris crassifolia voucher Douglas 2203 clone 1 phosphoenolpyruvate carboxylase (ppc-1E1) gene, partial cds	*Cyphomeris crassifolia*	86.89	6.44E-09
9325295_SilicoDArT	69	1009113634	XM_016017767	PREDICTED: Ziziphus jujuba zinc finger CCHC domain-containing protein 7-like (LOC107410354), mRNA	*Ziziphus jujuba*	100.00	6.48E-04
9325363_SilicoDArT	69	828339433	XM_004516695	PREDICTED: Cicer arietinum isoliquiritigenin 2'-O-methyltransferase-like (LOC101512403), mRNA	*Cicer arietinum*	100.00	1.80E-04
9325878_SilicoDArT	54	1040860766	XM_017385512	PREDICTED: Daucus carota subsp. sativus ethylene-responsive transcription factor ERF017-like (LOC108213718), mRNA	*Daucus carota subsp. sativus*	100.00	1.80E-04
9326946_SilicoDArT	69	1052182462	XM_017843383	PREDICTED: Phoenix dactylifera farnesyl pyrophosphate synthase 1-like (LOC103709296), transcript variant X2, mRNA	*Phoenix dactylifera*	94.74	3.87E-06
9328259_SilicoDArT	51	723218755	KF957690	Lycium barbarum P450 carotenoid beta-ring hydroxylase (CYP97A) mRNA, complete cds	*Lycium barbarum*	94.87	1.08E-06
9328493_SilicoDArT	69	743909762	XM_011050069	PREDICTED: Populus euphratica transcription factor bHLH68 (LOC105142436), transcript variant X2, mRNA	*Populus euphratica*	91.84	2.32E-08
9330263_SilicoDArT	54	1091481656	XM_010063113	PREDICTED: Eucalyptus grandis fructose-bisphosphate aldolase, cytoplasmic isozyme 1 (LOC104449094), mRNA	*Eucalyptus grandis*	88.37	6.48E-04

9330466_SilicoDArT	69	1012231113	XM_016082921	PREDICTED: Arachis duranensis heat shock 70 kDa protein 15-like (LOC107464023), mRNA	*Arachis duranensis*	92.86	8.27E-13
9330794_SilicoDArT	69	1040852357	XR_001804896	PREDICTED: Daucus carota subsp. sativus protein trichome berefringence-like 7 (LOC108210728), transcript variant X2, misc_RNA	*Daucus carota subsp. sativus*	92.11	1.80E-04
9330977_SilicoDArT	69	24460071	AB094592	Allium chinense lfs mRNA for lachrymatory factor synthase, complete cds	*Allium chinense*	94.20	4.91E-20
9330984_SilicoDArT	64	1091693398	XM_009400067	PREDICTED: Musa acuminata subsp. malaccensis serine/threonine-protein phosphatase BSL3-like (LOC103982958), transcript variant X6, mRNA	*Musa acuminata subsp. malaccensis*	92.19	1.37E-15
9331030_SilicoDArT	69	514747271	XM_004961269	PREDICTED: Setaria italica AP2/ERF and B3 domain-containing protein Os05g0549800 (LOC101782792), mRNA	*Setaria italica*	96.43	3.82E-16
9331040_SilicoDArT	69	823179024	XM_012631889	PREDICTED: Gossypium raimondii DNA mismatch repair protein MSH1, mitochondrial (LOC105800651), mRNA	*Gossypium raimondii*	100.00	1.08E-06
9331175_SilicoDArT	69	1022945697	KU318712	Allium cepa mitochondrion, complete genome	*Allium cepa*	100.00	1.05E-26
9331630_SilicoDArT	69	1121551646	JQ678773	Carica papaya GTP-binding nuclear protein (RAN) mRNA, partial cds	*Carica papaya*	100.00	1.06E-16
9331630_SilicoDArT	69	409058378	JX954153	Leea coccinea gene marker 1314 genomic sequence	*Leea coccinea*	93.22	1.78E-14
9332321_SilicoDArT	36	469402947	AB747100	Allium cepa AcRAD21-1 gene for cohesin subunit RAD21-1, partial cds, cultivar: Cheonjudaego	*Allium cepa*	100.00	2.32E-08
9333017_SilicoDArT	55	469402927	AB747098	Allium cepa AcRAD21-1 mRNA for cohesin subunit RAD21-1, complete cds, cultivar: Cheonjudaego	*Allium cepa*	98.04	4.94E-15
9333667_SilicoDArT	69	764534897	XM_004290660	PREDICTED: Fragaria vesca subsp. vesca aldo-keto reductase-like (LOC101305713), mRNA	*Fragaria vesca subsp. vesca*	100.00	6.48E-04
9333667_SilicoDArT	69	1039909549	XM_008351855	PREDICTED: Malus x domestica protein tas-like (LOC103413391), mRNA	*Malus domestica*	94.29	1.80E-04
9333976_SilicoDArT	69	1108943899	XM_010684266	PREDICTED: Beta vulgaris subsp. vulgaris RNA-binding protein 25 (LOC104897399), transcript variant X3, mRNA	*Beta vulgaris subsp. vulgaris*	100.00	6.48E-04
9333976_SilicoDArT	69	1109085562	XM_010478540	PREDICTED: Camelina sativa septin and tuftelin-interacting protein 1 homolog 1-like (LOC104756033), mRNA	*Camelina sativa*	100.00	6.48E-04
9333976_SilicoDArT	69	955337471	XM_003530589	PREDICTED: Glycine max F-box/kelch-repeat protein At1g23390 (LOC100799807), mRNA	*Glycine max*	100.00	6.48E-04
9333976_SilicoDArT	69	955380309	XM_003550331	PREDICTED: Glycine max S-adenosyl-L-methionine-dependent tRNA 4-demethylwyosine synthase-like (LOC100791920), mRNA	*Glycine max*	100.00	6.48E-04
9333976_SilicoDArT	69	1052171820	XM_008783796	PREDICTED: Phoenix dactylifera protein LATERAL ROOT PRIMORDIUM 1 (LOC103701650), transcript variant X2, mRNA	*Phoenix dactylifera*	100.00	6.48E-04
9333976_SilicoDArT	69	339409264	AC244896	Solanum lycopersicum strain Heinz 1706 chromosome 10 clone sle-7g18 map 10, complete sequence	*Solanum lycopersicum*	100.00	6.48E-04
9334504_SilicoDArT	69	404661	L12173	Allium porrum mannose specific lectin mRNA, complete cds	*Allium ampeloprasum*	100.00	1.80E-04
9334524_SilicoDArT	69	1050627799	XM_017781332	PREDICTED: Gossypium arboreum protein FEZ-like (LOC108478875), mRNA	*Gossypium arboreum*	100.00	1.80E-04
9334524_SilicoDArT	69	1083908480	XM_009631230	PREDICTED: Nicotiana tomentosiformis putative NAC domain-containing protein 94 (LOC104119659), transcript variant X2, mRNA	*Nicotiana tomentosiformis*	90.70	1.39E-05
9334524_SilicoDArT	69	1050594794	XM_017763371	PREDICTED: Gossypium arboreum transcription factor JUNGBRUNNEN 1-like (LOC108463429), mRNA	*Gossypium arboreum*	90.63	2.30E-13
9334748_SilicoDArT	69	802628256	XM_012221632	PREDICTED: Jatropha curcas 40S ribosomal protein S3-3-like (LOC105637942), mRNA	*Jatropha curcas*	100.00	6.48E-04

9335110_SilicoDArT	69	1104680471	XM_010666308	PREDICTED: Vitis vinifera mitogen-activated protein kinase YODA (LOC100257467), transcript variant X3, mRNA	*Vitis vinifera*	91.11	1.08E-06
9335547_SilicoDArT	69	1009113458	XM_016017670	PREDICTED: Ziziphus jujuba zinc finger CCHC domain-containing protein 7-like (LOC107410260), mRNA	*Ziziphus jujuba*	100.00	6.48E-04
9336023_SilicoDArT	69	1072932065	XM_018621197	PREDICTED: Raphanus sativus tubulin alpha-3 chain (LOC108847844), mRNA	*Raphanus sativus*	100.00	1.80E-04
9336238_SilicoDArT	69	848878926	XM_012984415	PREDICTED: Erythranthe guttatus APO protein 2, chloroplastic (LOC105960248), mRNA	*Mimulus guttatus*	100.00	6.48E-04
9336666_SilicoDArT	69	1028952386	XM_016864115	PREDICTED: Gossypium hirsutum sodium/hydrogen exchanger 2-like (LOC107932164), transcript variant X2, mRNA	*Gossypium hirsutum*	94.87	3.87E-06
9336922_SilicoDArT	69	1098789479	XM_019000533	PREDICTED: Juglans regia serine/threonine-protein phosphatase BSL3-like (LOC109018379), mRNA	*Juglans regia*	100.00	1.08E-06
9337242_SilicoDArT	69	282767699	GU253298	Allium cepa cultivar MRSPA ATP synthase subunit 6 (atp6) gene, complete cds; and unknown gene; mitochondrial	*Allium cepa*	100.00	1.39E-05
9337242_SilicoDArT	69	1022945697	KU318712	Allium cepa mitochondrion, complete genome	*Allium cepa*	100.00	6.48E-04
9338024_SilicoDArT	69	970037987	XM_015224844	PREDICTED: Solanum pennellii probable beta-1,3-galactosyltransferase 2 (LOC107023985), transcript variant X2, mRNA	*Solanum pennellii*	95.24	2.32E-08
9338065_SilicoDArT	69	1052171590	XM_008783501	PREDICTED: Phoenix dactylifera L-arabinokinase-like (LOC103701454), mRNA	*Phoenix dactylifera*	95.56	4.98E-10
9338202_SilicoDArT	69	955304399	XM_003517552	PREDICTED: Glycine max 40S ribosomal protein S29 (LOC100775561), mRNA	*Glycine max*	92.98	2.30E-13
9338212_SilicoDArT	69	469402936	AB747099	Allium cepa AcRAD21-1 gene for cohesin subunit RAD21-1, partial cds, cultivar: Cheonjudaego	*Allium cepa*	100.00	1.05E-26
9338332_SilicoDArT	69	1083885722	XM_009616079	PREDICTED: Nicotiana tomentosiformis ammonium transporter 3 member 1-like (LOC104107316), mRNA	*Nicotiana tomentosiformis*	89.86	1.78E-14
9339167_SilicoDArT	53	1052192815	XM_008805371	PREDICTED: Phoenix dactylifera AP2-like ethylene-responsive transcription factor AIL5 (LOC103717110), mRNA	*Phoenix dactylifera*	97.44	2.32E-08
9340282_SilicoDArT	40	1012205831	XM_016076375	PREDICTED: Arachis duranensis sedoheptulose-1,7-bisphosphatase, chloroplastic (LOC107458166), mRNA	*Arachis duranensis*	94.60	1.39E-05
9340426_SilicoDArT	32	469402947	AB747100	Allium cepa AcRAD21-1 gene for cohesin subunit RAD21-1, partial cds, cultivar: Cheonjudaego	*Allium cepa*	100.00	3.87E-06
9340864_SilicoDArT	48	49781342	AY647262	Allium cepa flavonol synthase gene, complete cds	*Allium cepa*	100.00	4.94E-15
9341417_SilicoDArT	69	828303965	XM_012714488	PREDICTED: Cicer arietinum glycogen synthase kinase-3 homolog MsK-1 (LOC101500512), transcript variant X3, mRNA	*Cicer arietinum*	89.13	5.01E-05
9341988_SilicoDArT	69	1130672164	XM_019852734	PREDICTED: Elaeis guineensis glucose-6-phosphate 1-dehydrogenase, chloroplastic-like (LOC105051009), transcript variant X2, mRNA	*Elaeis guineensis*	87.50	1.38E-10
9341999_SilicoDArT	69	166342	M94106	Allium sativum chitinase mRNA, 3' end	*Allium sativum*	100.00	1.05E-26
9341999_SilicoDArT	69	166342	M94106	Allium sativum chitinase mRNA, 3' end	*Allium sativum*	100.00	1.05E-26
9342092_SilicoDArT	69	747087535	XM_011093012	PREDICTED: Sesamum indicum malate dehydrogenase (LOC105171782), mRNA	*Sesamum indicum*	88.06	2.97E-12
9342221_SilicoDArT	69	1050573319	XM_017752012	PREDICTED: Gossypium arboreum sterol 3-beta-glucosyltransferase UGT80B1 (LOC108453732), transcript variant X4, mRNA	*Gossypium arboreum*	96.08	2.30E-13
9342791_SilicoDArT	69	848884676	XM_012987148	PREDICTED: Erythranthe guttatus uncharacterized LOC105962815 (LOC105962815), transcript variant X3, mRNA	*Mimulus guttatus*	94.00	3.85E-11

9343366_SilicoDArT	69	927028619	AB924383	Allium sativum AsFMO1 mRNA for S-allyl-L-cysteine S-oxygenase, complete cds	*Allium sativum*	92.86	2.28E-18
9343366_SilicoDArT	69	927028619	AB924383	Allium sativum AsFMO1 mRNA for S-allyl-L-cysteine S-oxygenase, complete cds	*Allium sativum*	92.86	2.28E-18
9344037_SilicoDArT	69	780981384	KM117265	Allium cepa 18S ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, internal transcribed spacer 2, and 26S ribosomal RNA gene, complete sequence	*Allium cepa*	97.10	2.27E-23
9344037_SilicoDArT	69	15407260	AY034663	Peltoboykinia tellimoides large subunit ribosomal RNA gene, partial sequence	*Peltoboykinia tellimoides*	95.65	1.05E-21
9344094_SilicoDArT	69	828303334	XM_004495312	PREDICTED: Cicer arietinum serine/threonine protein phosphatase 2A regulatory subunit B"beta-like (LOC101505006), mRNA	*Cicer arietinum*	92.50	1.39E-05
9344094_SilicoDArT	69	1040902159	XM_017403219	PREDICTED: Daucus carota subsp. sativus serine/threonine protein phosphatase 2A regulatory subunit B"beta-like (LOC108227854), mRNA	*Daucus carota subsp. sativus*	97.50	6.44E-09
9344094_SilicoDArT	69	848884423	XM_012987028	PREDICTED: Erythranthe guttatus serine/threonine protein phosphatase 2A regulatory subunit B"beta-like (LOC105962710), mRNA	*Mimulus guttatus*	97.50	6.44E-09
9344094_SilicoDArT	69	747061496	XM_011078918	PREDICTED: Sesamum indicum serine/threonine protein phosphatase 2A regulatory subunit B"beta-like (LOC105161279), mRNA	*Sesamum indicum*	95.00	2.99E-07
9344172_SilicoDArT	69	1081749611	KT935444	Allium cepa fructan: fructan 6G-fructosyltransferase mRNA, complete cds	*Allium cepa*	100.00	1.05E-26
9344888_SilicoDArT	44	1040902836	XM_017403534	PREDICTED: Daucus carota subsp. sativus F-box protein PP2-A15 (LOC108228058), mRNA	*Daucus carota subsp. sativus*	100.00	5.01E-05
9345409_SilicoDArT	59	148872676	EF633511	Allium cepa var. aggregatum lipid transfer protein 4 gene, complete cds	*Allium ascalonicum*	100.00	3.79E-21
9345409_SilicoDArT	59	171221510	EU561064	Allium cepa antimicrobial peptide mRNA, partial cds	*Allium cepa*	100.00	3.79E-21
9345778_SilicoDArT	69	1083874822	XM_009608522	PREDICTED: Nicotiana tomentosiformis DUF21 domain-containing protein At4g14240-like (LOC104101106), mRNA	*Nicotiana tomentosiformis*	85.71	2.32E-08
9345846_SilicoDArT	69	1040916157	XM_017364840	PREDICTED: Daucus carota subsp. sativus uncharacterized protein At1g04910-like (LOC108197265), mRNA	*Daucus carota subsp. sativus*	97.06	1.39E-05
9345909_SilicoDArT	69	1026013575	XM_016711254	PREDICTED: Capsicum annuum phosphopantothenoylcysteine decarboxylase-like (LOC107864830), transcript variant X2, mRNA	*Capsicum annuum*	88.53	4.98E-10
9346281_SilicoDArT	69	1002841322	XM_015833092	PREDICTED: Oryza brachyantha thaumatin-like protein 1b (LOC102716100), mRNA	*Oryza brachyantha*	84.13	3.87E-06
9346529_SilicoDArT	69	778662940	XM_011661678	PREDICTED: Cucumis sativus beta-galactosidase-like (LOC101218515), mRNA	*Cucumis sativus*	94.60	1.39E-05
9346791_SilicoDArT	69	1109103029	XM_010494896	PREDICTED: Camelina sativa probable pectinesterase/pectinesterase inhibitor 54 (LOC104770464), mRNA	*Camelina sativa*	100.00	1.38E-10
9347750_SilicoDArT	44	1035395921	XM_008441249	PREDICTED: Cucumis melo protein CHROMATIN REMODELING 4 (LOC103484261), transcript variant X4, mRNA	*Cucumis melo*	88.64	1.80E-04
9348548_SilicoDArT	69	1130632797	XM_010938880	PREDICTED: Elaeis guineensis chaperonin-like RbcX protein 2, chloroplastic (LOC105056619), mRNA	*Elaeis guineensis*	93.85	8.21E-18
9348634_SilicoDArT	69	1091648932	XM_009423084	PREDICTED: Musa acuminata subsp. malaccensis calcium-transporting ATPase 1, endoplasmic reticulum-type-like (LOC104000928), mRNA	*Musa acuminata subsp. malaccensis*	90.63	6.39E-14
9349862_SilicoDArT	69	793420965	NM_001305731	Cucumis sativus probable sucrose-phosphate synthase 2 (LOC101208942), mRNA	*Cucumis sativus*	96.97	5.01E-05
9349875_SilicoDArT	61	1109238907	XM_019338588	PREDICTED: Ipomoea nil phosphoenolpyruvate carboxylase-like (LOC109188085), transcript variant X1, mRNA	*Ipomoea nil*	93.48	6.44E-09

9350048_SilicoDArT	61	1063488637	XM_007032508	PREDICTED: Theobroma cacao polypyrimidine tract-binding protein homolog 1 (LOC18601541), mRNA	*Theobroma cacao*	90.91	1.38E-10
9350049_SilicoDArT	61	1091726857	XM_009405808	PREDICTED: Musa acuminata subsp. malaccensis polypyrimidine tract-binding protein homolog 1-like (LOC103987487), transcript variant X3, mRNA	*Musa acuminata subsp. malaccensis*	94.44	2.30E-13
9351005_SilicoDArT	69	923912599	XM_013870530	PREDICTED: Brassica napus GDSL esterase/lipase LTL1-like (LOC106429781), mRNA	*Brassica napus*	95.46	1.79E-09
9351041_SilicoDArT	69	596187048	XM_007223302	Prunus persica hypothetical protein (PRUPE_ppa005616mg) mRNA, complete cds	*Prunus persica*	97.14	3.87E-06
9351221_SilicoDArT	69	1108935616	XM_010681853	PREDICTED: Beta vulgaris subsp. vulgaris probable protein S-acyltransferase 19 (LOC104895372), mRNA	*Beta vulgaris subsp. vulgaris*	94.12	1.76E-19
9351668_SilicoDArT	53	1126731859	LT669794	Arabis alpina genome assembly, chromosome: chr7	*Arabis alpina*	100.00	1.80E-04
9351668_SilicoDArT	53	1126731459	LT669789	Arabis alpina genome assembly, chromosome: chr2	*Arabis alpina*	100.00	1.80E-04
9351778_SilicoDArT	35	1025307851	XM_016645574	PREDICTED: Nicotiana tabacum transcription factor MYB26-like (LOC107819474), transcript variant X2, mRNA	*Nicotiana tabacum*	100.00	6.48E-04
9352361_SilicoDArT	69	700253107	KM397511	Asparagus officinalis MSH1 mRNA, partial cds	*Asparagus officinalis*	93.62	6.44E-09
9352365_SilicoDArT	69	1027091046	XM_008232361	PREDICTED: Prunus mume E3 ubiquitin-protein ligase RHF2A (LOC103329838), mRNA	*Prunus mume*	97.62	4.98E-10
9352474_SilicoDArT	69	1091484874	XM_010064534	PREDICTED: Eucalyptus grandis protein transport protein Sec61 subunit alpha (LOC104450102), mRNA	*Eucalyptus grandis*	94.60	1.39E-05
9353093_SilicoDArT	60	848861233	XM_012975986	PREDICTED: Erythranthe guttatus 3-hydroxyisobutyryl-CoA hydrolase 1-like (LOC105952435), mRNA	*Mimulus guttatus*	92.11	1.80E-04
9353451_SilicoDArT	69	510122042	KC466030	Allium cepa UFGT2 mRNA, complete cds	*Allium cepa*	100.00	1.05E-26
9353780_SilicoDArT	69	297809984	XM_002872830	Arabidopsis lyrata subsp. lyrata protein translocase subunit secA chloroplast precursor, mRNA	*Arabidopsis lyrata subsp. lyrata*	92.06	4.94E-15
9353802_SilicoDArT	69	1052178563	XM_008790719	PREDICTED: Phoenix dactylifera zinc finger protein CONSTANS-like (LOC103706571), transcript variant X4, mRNA	*Phoenix dactylifera*	85.94	6.44E-09
9353853_SilicoDArT	69	1102790753	XM_010245323	PREDICTED: Nelumbo nucifera myosin-6-like (LOC104587632), mRNA	*Nelumbo nucifera*	94.29	1.80E-04
9354458_SilicoDArT	69	1098721498	XM_018989860	PREDICTED: Juglans regia casein kinase 1-like protein 1 (LOC109009397), mRNA	*Juglans regia*	89.23	8.27E-13
9354508_SilicoDArT	63	747066802	XM_011081795	PREDICTED: Sesamum indicum cellulose synthase A catalytic subunit 4 [UDP-forming] (LOC105163451), mRNA	*Sesamum indicum*	89.83	3.85E-11
9354557_SilicoDArT	69	1130635044	XM_010906974	PREDICTED: Elaeis guineensis BTB/POZ domain-containing protein At5g03250-like (LOC105032513), mRNA	*Elaeis guineensis*	97.22	1.08E-06
9354568_SilicoDArT	69	42565431	AY389732	Hyacinthus orientalis 40S ribosomal protein S23 mRNA, partial cds	*Hyacinthus orientalis*	98.11	3.82E-16
9354630_SilicoDArT	69	242041800	XM_002468250	Sorghum bicolor hypothetical protein, mRNA	*Sorghum bicolor*	94.44	1.80E-04
9354789_SilicoDArT	65	1098803480	XM_018952042	PREDICTED: Juglans regia mitogen-activated protein kinase 8-like (LOC108980971), mRNA	*Juglans regia*	92.31	3.82E-16
9354847_SilicoDArT	69	1098820251	XM_018960766	PREDICTED: Juglans regia RNA polymerase II-associated protein 3 (LOC108987762), transcript variant X3, mRNA	*Juglans regia*	88.41	2.30E-13

9354897_SilicoDArT	69	960458686	XM_014897296	PREDICTED: Brachypodium distachyon ribulose-phosphate 3-epimerase, chloroplastic (LOC100826802), transcript variant X2, mRNA	*Brachypodium distachyon*	95.35	6.44E-09
9355046_SilicoDArT	60	1130627979	XM_010910495	PREDICTED: Elaeis guineensis galacturonosyltransferase 8-like (LOC105035085), mRNA	*Elaeis guineensis*	88.14	1.79E-09
9355352_SilicoDArT	69	358248409	NM_001252692	Glycine max 40S ribosomal protein S3-3-like (LOC100808705), mRNA	*Glycine max*	85.51	4.98E-10
9355596_SilicoDArT	69	1105486800	XM_010650786	PREDICTED: Vitis vinifera beta-galactosidase 3 (LOC100232848), mRNA	*Vitis vinifera*	88.14	1.79E-09
9355597_SilicoDArT	69	922534331	XM_013742955	PREDICTED: Brassica oleracea var. oleracea beta-galactosidase 3-like (LOC106306365), mRNA	*Brassica oleracea var. oleracea*	97.14	3.87E-06
9356271_SilicoDArT	69	1021570407	XM_016320901	PREDICTED: Arachis ipaensis peroxidase A2-like (LOC107618757), mRNA	*Arachis ipaensis*	97.06	1.39E-05
9356301_SilicoDArT	69	769794621	XM_011630423	PREDICTED: Amborella trichopoda BTB/POZ and TAZ domain-containing protein 3 (LOC18448701), transcript variant X3, mRNA	*Amborella trichopoda*	95.46	1.79E-09
9356799_SilicoDArT	69	296184580	GU573766	Dimocarpus longan cultivar Honghezi 14-3-3 family protein gene, complete cds	*Dimocarpus longan*	96.88	1.80E-04
9356799_SilicoDArT	69	218202931	FJ479618	Dimocarpus longan 14-3-3 protein mRNA, complete cds	*Dimocarpus longan*	96.88	1.80E-04
9356800_SilicoDArT	69	1040831491	XM_017374550	PREDICTED: Daucus carota subsp. sativus 14-3-3-like protein (LOC108204889), mRNA	*Daucus carota subsp. sativus*	100.00	1.80E-04
9356900_SilicoDArT	69	1052182462	XM_017843383	PREDICTED: Phoenix dactylifera farnesyl pyrophosphate synthase 1-like (LOC103709296), transcript variant X2, mRNA	*Phoenix dactylifera*	97.22	1.08E-06
9357227_SilicoDArT	60	1000986181	XM_002510541	PREDICTED: Ricinus communis peroxidase 64 (LOC8267824), mRNA	*Ricinus communis*	90.00	1.07E-11
9357457_SilicoDArT	69	922467943	XM_013778634	PREDICTED: Brassica oleracea var. oleracea protein transport protein Sec61 subunit beta-like (LOC106339764), mRNA	*Brassica oleracea var. oleracea*	100.00	1.39E-05
9357544_SilicoDArT	69	778682834	XM_011653489	PREDICTED: Cucumis sativus clustered mitochondria protein (LOC101207687), transcript variant X1, mRNA	*Cucumis sativus*	96.97	5.01E-05
9358018_SilicoDArT	69	1050591662	XM_017761681	PREDICTED: Gossypium arboreum protein mago nashi homolog (LOC108461752), mRNA	*Gossypium arboreum*	93.33	4.94E-15
9358018_SilicoDArT	69	1063871953	XM_018145565	Phialophora attae hypothetical protein mRNA	*Phialophora attae*	97.30	2.99E-07
9358259_SilicoDArT	69	703088396	XM_010095222	Morus notabilis Germination-specific cysteine protease 1 partial mRNA	*Morus notabilis*	96.97	5.01E-05
9358310_SilicoDArT	69	1052183060	XM_008795256	PREDICTED: Phoenix dactylifera dehydrogenase/reductase SDR family member 12 (LOC103709761), mRNA	*Phoenix dactylifera*	89.23	2.97E-12
9358363_SilicoDArT	51	836002706	XM_004976714	PREDICTED: Setaria italica 26S proteasome non-ATPase regulatory subunit 8 homolog A-like (LOC101769553), mRNA	*Setaria italica*	97.44	2.32E-08
9359097_SilicoDArT	69	1111112471	XM_019392296	PREDICTED: Nicotiana attenuata dnaJ protein ERDJ3B-like (LOC109227226), mRNA	*Nicotiana attenuata*	92.86	8.27E-13
9360464_SilicoDArT	43	741985377	KM434203	Allium cepa dihydroflavonol 4-reductase (DFR-A) gene, DFR-ADTP allele, complete cds; and transposon AcCACTA1, complete sequence	*Allium cepa*	100.00	3.85E-11
9360611_SilicoDArT	60	662247390	KJ719265	Nicotiana tabacum HSP18.9 mRNA, complete cds	*Nicotiana tabacum*	100.00	2.99E-07

9360611_SilicoDArT	60	99033682	DQ515772	Agave tequilana clone 1.8 chaperone mRNA, complete cds	*Agave tequilana*	100.00	6.48E-04
9360611_SilicoDArT	60	326528088	AK357882	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone: NIASHv1064A22	*Hordeum vulgare subsp. vulgare*	93.02	2.99E-07
9360612_SilicoDArT	60	99033708	DQ515785	Agave tequilana clone 7.8 chaperone mRNA, complete cds	*Agave tequilana*	97.14	3.87E-06
9360612_SilicoDArT	60	326504765	AK375479	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone: NIASHv3095H06	*Hordeum vulgare subsp. vulgare*	92.68	1.39E-05
9360612_SilicoDArT	60	1109265466	XM_019301333	PREDICTED: Ipomoea nil 18.2 kDa class I heat shock protein-like (LOC109153464), mRNA	*Ipomoea nil*	92.68	1.39E-05
9360913_SilicoDArT	37	469402947	AB747100	Allium cepa AcRAD21-1 gene for cohesin subunit RAD21-1, partial cds, cultivar: Cheonjudaego	*Allium cepa*	100.00	6.44E-09

Supplementary Material 2.2. UniProt analyses of DArTseq reads. ID: DArTseq identification number; GI: gene identification; Description: NCBI gene description; PIC: polymorphism-information content; UNIPROT ID: protein identification (for UniProt database); ATH ID: *Arabidopsis thaliana* ID in UniProt database; GO: Gene Ontology; GO name: GO-term explanation; GO domain: Biological Process (BP); Cellular Component (CC); Molecular Function (MF).

ID	GI	Description	PIC	UNIPROT ID	ATH ID	GO	GO name	GO domain
9321425_SilicoDArT	1109138320	PREDICTED: Camelina sativa 29 kDa ribonucleoprotein, chloroplastic (LOC104791610), mRNA	0.22	Q43349	Q43349	GO:0000166	Nucleotide binding	MF
						GO:0003676	Nucleic acid binding	MF
						GO:0003723	RNA binding	MF
						GO:0006397	mRNA processing	BP
						GO:0009507	Chloroplast	CC
						GO:0009536	Plastid	CC
						GO:0009631	Cold acclimation	BP
						GO:0030529	Intracellular ribonucleoprotein complex	CC
						GO:0043489	RNA stabilization	BP
9321555_SilicoDArT	1026026375	PREDICTED: Capsicum annuum 40S ribosomal protein S4-like (LOC107859079), mRNA	0.26	K4B818	A0A178UQ69	GO:0003723	RNA binding	MF
						GO:0003735	Structural constituent of ribosome	MF
						GO:0005622	Intracellular	CC
						GO:0005840	Ribosome	CC
						GO:0006412	Translation	BP
						GO:0019843	rRNA binding	MF
						GO:0022627	Cytosolic small ribosomal subunit	CC
						GO:0030529	Intracellular ribonucleoprotein complex	CC
9322412_SilicoDArT	59799342	Allium sativum chloroplast cysteine synthase GCS2 (gcs2) mRNA, complete cds; nuclear gene for chloroplast product	0.21	Q3L197	Q0WW95	GO:0004124	Cysteine synthase activity	MF
						GO:0006535	Cysteine biosynthetic process from serine	BP
						GO:0008652	Cellular amino acid biosynthetic process	BP
						GO:0016740	Transferase activity	MF
						GO:0019344	Cysteine biosynthetic process	BP

9322838_SilicoDArT	828319638	PREDICTED: Cicer arietinum rac-like GTP-binding protein RAC13 (LOC101498092), mRNA	0.18	B6CHW8	Q38903	GO:0005525	GTP binding	MF
						GO:0005622	Intracellular	CC
						GO:0007264	Small GTPase mediated signal transduction	BP
9323004_SilicoDArT	27448223	Allium sativum phytochelatins synthase (pcs1) mRNA, complete cds	0.47	Q8GZS8	Q9S7Z3	GO:0010038	Response to metal ion	BP
						GO:0016756	Glutathione gamma-glutamylcysteinyltransferase activity	MF
						GO:0046872	Metal ion binding	MF
						GO:0046938	Phytochelatin biosynthetic process	BP
9323284_SilicoDArT	1117415913	PREDICTED: Lupinus angustifolius auxin transporter-like protein 5 (LOC109325971), transcript variant X3, mRNA	0.27	V7ADR7	Q9S836	GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
9323367_SilicoDArT	1102790753	PREDICTED: Nelumbo nucifera myosin-6-like (LOC104587632), mRNA	0.41	A0A061DN86	Q9LKB9	GO:0000166	Nucleotide binding	MF
						GO:0003774	Motor activity	MF
						GO:0003779	Actin binding	MF
						GO:0005524	ATP binding	MF
						GO:0016459	Myosin complex	CC
9323600_SilicoDArT	662247392	Nicotiana tabacum HSP19.8 mRNA, partial cds	0.25	P27879		GO:0005737	Cytoplasm	CC
9324289_SilicoDArT	802555686	PREDICTED: Jatropha curcas polygalacturonase-like (LOC105628710), mRNA	0.35	A0A067L4X9	Q9SFB7	GO:0004650	Polygalacturonase activity	MF
						GO:0005576	Extracellular region	CC
						GO:0005975	Carbohydrate metabolic process	BP
						GO:0008152	Metabolic process	BP
						GO:0016787	Hydrolase activity	MF
						GO:0016798	Hydrolase activity, acting on glycosyl bonds	MF
						GO:0071555	Cell wall organization	BP
9325222_SilicoDArT	330689878	Allium roylei lachrymatory factor synthase (LFS) gene, partial cds	0.32	P59082		GO:0005773	Vacuole	CC
9325282_SilicoDArT	664680186	Cyphomeris crassifolia voucher Douglas 2203 clone 1 phosphoenolpyruvate carboxylase (ppc-1E1) gene, partial cds	0.32	A0A075J595		GO:0003824	Catalytic activity	MF
						GO:0006099	Tricarboxylic acid cycle	BP

						GO:0008964	Phosphoenolpyruvate carboxylase activity	MF
						GO:0015977	Carbon fixation	BP
9325295_SilicoDArT	1009113634	PREDICTED: Ziziphus jujuba zinc finger CCHC domain-containing protein 7-like (LOC107410354), mRNA	0.38	M5WL34	Q9FG62	GO:0003676	Nucleic acid binding	MF
						GO:0008270	Zinc ion binding	MF
9325363_SilicoDArT	828339433	PREDICTED: Cicer arietinum isoliquiritigenin 2'-O-methyltransferase-like (LOC101512403), mRNA	0.32	A0A072VH10	Q9FK25	GO:0008168	Methyltransferase activity	MF
						GO:0008171	O-methyltransferase activity	MF
						GO:0016740	Transferase activity	MF
						GO:0032259	Methylation	BP
						GO:0046983	Protein dimerization activity	MF
9325878_SilicoDArT	1040860766	PREDICTED: Daucus carota subsp. sativus ethylene-responsive transcription factor ERF017-like (LOC108213718), mRNA	0.31	A0A166C1Z0	Q84QC2	GO:0003677	DNA binding	MF
						GO:0003700	Transcription factor activity, sequence-specific DNA binding	MF
						GO:0005634	Nucleus	CC
						GO:0006351	Transcription, DNA-templated	BP
						GO:0006355	Regulation of transcription, DNA-templated	BP
9326946_SilicoDArT	1052182462	PREDICTED: Phoenix dactylifera farnesyl pyrophosphate synthase 1-like (LOC103709296), transcript variant X2, mRNA	0.37	M0SIL9	Q09152	GO:0008299	Isoprenoid biosynthetic process	BP
						GO:0016740	Transferase activity	MF
9328259_SilicoDArT	723218755	Lycium barbarum P450 carotenoid beta-ring hydroxylase (CYP97A) mRNA, complete cds	0.10	A0A0A7DVQ0	Q6TBX7	GO:0004497	Monoxygenase activity	MF
						GO:0005506	Iron ion binding	MF
						GO:0016491	Oxidoreductase activity	MF
						GO:0016705	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	MF
						GO:0020037	Heme binding	MF
						GO:0046872	Metal ion binding	MF
						GO:0055114	Oxidation-reduction process	BP
9328493_SilicoDArT	743909762	PREDICTED: Populus euphratica transcription factor bHLH68 (LOC105142436), transcript variant X2, mRNA	0.02	B9ILQ2	Q8S3D1	GO:0001046	Core promoter sequence-specific DNA binding	MF

						GO:0001228	Transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding	MF
						GO:0005634	Nucleus	CC
						GO:0006366	Transcription from RNA polymerase II promoter	BP
						GO:0045944	Positive regulation of transcription from RNA polymerase II promoter	BP
						GO:0046983	Protein dimerization activity	MF
9330263_SilicoDArT	1091481656	PREDICTED: Eucalyptus grandis fructose-bisphosphate aldolase, cytoplasmic isozyme 1 (LOC104449094), mRNA	0.00	A0A059BQ87	O65581	GO:0003824	Catalytic activity	MF
						GO:0004332	Fructose-bisphosphate aldolase activity	MF
						GO:0006096	Glycolytic process	BP
						GO:0016829	Lyase activity	MF
9330466_SilicoDArT	1012231113	PREDICTED: Arachis duranensis heat shock 70 kDa protein 15-like (LOC107464023), mRNA	0.00	V7BXL6	Q9S7C0	GO:0000166	Nucleotide binding	MF
						GO:0005524	ATP binding	MF
9330794_SilicoDArT	1040852357	PREDICTED: Daucus carota subsp. sativus protein trichome berefringence-like 7 (LOC108210728), transcript variant X2, misc_RNA	0.01	A0A166BNS6	F4I037	GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
9330977_SilicoDArT	24460071	Allium chinense lfs mRNA for lachrymatory factor synthase, complete cds	0.00	P59082		GO:0005773	Vacuole	CC
9330984_SilicoDArT	1091693398	PREDICTED: Musa acuminata subsp. malaccensis serine/threonine-protein phosphatase BSL3-like (LOC103982958), transcript variant X6, mRNA	0.01	M0SR93	Q9SHS7	GO:0004721	phosphoprotein phosphatase activity	MF
						GO:0006470	protein dephosphorylation	BP
						GO:0009742	brassinosteroid mediated signaling pathway	BP
						GO:0016787	hydrolase activity	MF
9331030_SilicoDArT	514747271	PREDICTED: Setaria italica AP2/ERF and B3 domain-containing protein Os05g0549800 (LOC101782792), mRNA	0.01	K3ZDE0	P82280	GO:0003677	DNA binding	MF
						GO:0003700	Transcription factor activity, sequence-specific DNA binding	MF
						GO:0005634	Nucleus	CC

						GO:0006351	Transcription, DNA-templated	BP
						GO:0006355	Regulation of transcription, DNA-templated	BP
9331040_SilicoDArT	823179024	PREDICTED: Gossypium raimondii DNA mismatch repair protein MSH1, mitochondrial (LOC105800651), mRNA	0.00	A0A0D2S9J8	Q84LK0	GO:0005524	ATP binding	MF
						GO:0006298	Mismatch repair	BP
						GO:0030983	Mismatched DNA binding	MF
9331630_SilicoDArT	1121551646	Carica papaya GTP-binding nuclear protein (RAN) mRNA, partial cds	0.00	M4I2U2	Q8H156	GO:0000166	Nucleotide binding	MF
						GO:0003924	GTPase activity	MF
						GO:0005525	GTP binding	MF
						GO:0005634	Nucleus	CC
						GO:0006810	Transport	BP
						GO:0006886	Intracellular protein transport	BP
						GO:0006913	Nucleocytoplasmic transport	BP
						GO:0007165	Signal transduction	BP
						GO:0007264	Small GTPase mediated signal transduction	BP
						GO:0015031	Protein transport	BP
9331630_SilicoDArT	409058378	Leea coccinea gene marker 1314 genomic sequence	0.00	R9Q9C5	Q8H156	GO:0000166	Nucleotide binding	MF
						GO:0003924	GTPase activity	MF
						GO:0005525	GTP binding	MF
						GO:0005634	Nucleus	CC
						GO:0006810	Transport	BP
						GO:0006886	Intracellular protein transport	BP
						GO:0006913	Nucleocytoplasmic transport	BP
						GO:0007165	Signal transduction	BP
						GO:0007264	Small GTPase mediated signal transduction	BP
						GO:0015031	Protein transport	BP
9332321_SilicoDArT	469402947	Allium cepa AcRAD21-1 gene for cohesin subunit RAD21-1, partial cds, cultivar: Cheonjudaego	0.00	M5A8H0	Q8W1Y0	GO:0000228	Nuclear chromosome	CC

9333667_SilicoDArT	764534897	PREDICTED: Fragaria vesca subsp. vesca aldo-keto reductase-like (LOC101305713), mRNA	0.00	NA		NA	NA	NA
9333976_SilicoDArT	1108943899	PREDICTED: Beta vulgaris subsp. vulgaris RNA-binding protein 25 (LOC104897399), transcript variant X3, mRNA	0.00	A0A0J8C0H4		GO:0000166	Nucleotide binding	MF
						GO:0003676	Nucleic acid binding	MF
						GO:0006397	mRNA processing	BP
9333976_SilicoDArT	1109085562	PREDICTED: Camelina sativa septin and tuftelin-interacting protein 1 homolog 1-like (LOC104756033), mRNA	0.00	R0GUE3	Q9SHG6	GO:0003676	Nucleic acid binding	MF
						GO:0003677	DNA binding	MF
						GO:0005634	Nucleus	CC
						GO:0005681	Spliceosomal complex	CC
						GO:0006355	Regulation of transcription, DNA-templated	BP
						GO:0006397	mRNA processing	BP
						GO:0008380	RNA splicing	BP
9333976_SilicoDArT	955337471	PREDICTED: Glycine max F-box/kelch-repeat protein At1g23390 (LOC100799807), mRNA	0.00	NA		NA	NA	NA
9333976_SilicoDArT	955380309	PREDICTED: Glycine max S-adenosyl-L-methionine-dependent tRNA 4-demethylwyosine synthase-like (LOC100791920), mRNA	0.00	NA		NA	NA	NA
9333976_SilicoDArT	1052171820	PREDICTED: Phoenix dactylifera protein LATERAL ROOT PRIMORDIUM 1 (LOC103701650), transcript variant X2, mRNA	0.00	Q94CK9	Q94CK9	GO:0003677	DNA binding	MF
						GO:0005634	Nucleus	CC
						GO:0007275	Multicellular organism development	BP
						GO:0009734	Auxin-activated signaling pathway	BP
						GO:0009851	Auxin biosynthetic process	BP
						GO:0042803	Protein homodimerization activity	MF
						GO:0046872	Metal ion binding	MF
9334504_SilicoDArT	404661	Allium porrum mannose specific lectin mRNA, complete cds	0.23	Q38759		GO:0030246	Carbohydrate binding	MF
9334524_SilicoDArT	1050627799	PREDICTED: Gossypium arboreum protein FEZ-like (LOC108478875), mRNA	0.20	A0A0D2SGR3	Q9ZVH0	GO:0003677	DNA binding	MF
						GO:0005634	Nucleus	CC
						GO:0006351	Transcription, DNA-templated	BP
						GO:0006355	Regulation of transcription, DNA-templated	BP

9334524_SilicoDArT	1083908480	PREDICTED: Nicotiana tomentosiformis putative NAC domain-containing protein 94 (LOC104119659), transcript variant X2, mRNA	0.20	M1D3A9	Q9FIW5	GO:0003677	DNA binding	MF
						GO:0005634	Nucleus	CC
						GO:0006351	Transcription, DNA-templated	BP
						GO:0006355	Regulation of transcription, DNA-templated	BP
9334524_SilicoDArT	1050594794	PREDICTED: Gossypium arboreum transcription factor JUNGBRUNNEN 1-like (LOC108463429), mRNA	0.20	A0A0D2SAQ4	Q9SK55	GO:0003677	DNA binding	MF
						GO:0005634	Nucleus	CC
						GO:0006351	Transcription, DNA-templated	BP
						GO:0006355	Regulation of transcription, DNA-templated	BP
9334748_SilicoDArT	802628256	PREDICTED: Jatropha curcas 40S ribosomal protein S3-3-like (LOC105637942), mRNA	0.23	A0A067KFX0	Q9FJA6	GO:0003723	RNA binding	MF
						GO:0003735	Structural constituent of ribosome	MF
						GO:0005840	Ribosome	CC
						GO:0006412	Translation	BP
						GO:0015935	Small ribosomal subunit	CC
						GO:0030529	Intracellular ribonucleoprotein complex	CC
9335110_SilicoDArT	1104680471	PREDICTED: Vitis vinifera mitogen-activated protein kinase YODA (LOC100257467), transcript variant X3, mRNA	0.19	F6H1E6	Q9C5H5	GO:0004672	Protein kinase activity	MF
						GO:0004702	Signal transducer, downstream of receptor, with serine/threonine kinase activity	MF
						GO:0005524	ATP binding	MF
						GO:0005737	Cytoplasm	CC
						GO:0006468	Protein phosphorylation	BP
						GO:0023014	Signal transduction by protein phosphorylation	BP
9335547_SilicoDArT	1009113458	PREDICTED: Ziziphus jujuba zinc finger CCHC domain-containing protein 7-like (LOC107410260), mRNA	0.45	M5WL34	Q9FG62	GO:0003676	Nucleic acid binding	MF
						GO:0008270	Zinc ion binding	MF
9336023_SilicoDArT	1072932065	PREDICTED: Raphanus sativus tubulin alpha-3 chain (LOC108847844), mRNA	0.25	A0A178UTQ9	Q56WH1	GO:0000166	Nucleotide binding	MF
						GO:0003924	GTPase activity	MF

						GO:0005200	Structural constituent of cytoskeleton	MF
						GO:0005525	GTP binding	MF
						GO:0005874	Microtubule	CC
						GO:0007010	Cytoskeleton organization	BP
						GO:0007017	Microtubule-based process	BP
9336238_SilicoDArT	848878926	PREDICTED: Erythranthe guttatus APO protein 2, chloroplastic (LOC105960248), mRNA	0.21	A0A022R8W3	Q8W4A5	GO:0003723	RNA binding	MF
9336666_SilicoDArT	1028952386	PREDICTED: Gossypium hirsutum sodium/hydrogen exchanger 2-like (LOC107932164), transcript variant X2, mRNA	0.23	A0A0D2PN73	Q56XP4	GO:0005774	Vacuolar membrane	CC
						GO:0005886	Plasma membrane	CC
						GO:0006810	Transport	BP
						GO:0006811	Ion transport	BP
						GO:0006812	Cation transport	BP
						GO:0006814	Sodium ion transport	BP
						GO:0006885	Regulation of pH	BP
						GO:0009651	Response to salt stress	BP
						GO:0015297	Antiporter activity	MF
						GO:0015299	Solute: proton antiporter activity	MF
						GO:0015385	Sodium: proton antiporter activity	MF
						GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
						GO:0035725	Sodium ion transmembrane transport	BP
						GO:0055075	Potassium ion homeostasis	BP
						GO:0055085	Transmembrane transport	BP
						GO:1902600	Hydrogen ion transmembrane transport	BP
9336922_SilicoDArT	1098789479	PREDICTED: Juglans regia serine/threonine-protein phosphatase BSL3-like (LOC109018379), mRNA	0.36	Q9SHS7	Q9SHS7	GO:0004721	Phosphoprotein phosphatase activity	MF
						GO:0005634	Nucleus	CC
						GO:0005829	Cytosol	CC

						GO:0005886	Plasma membrane	CC
						GO:0006470	Protein dephosphorylation	BP
						GO:0009742	Brassinosteroid mediated signaling pathway	BP
						GO:0016787	Hydrolase activity	MF
						GO:0046872	Metal ion binding	MF
						GO:0005829	Cytosol	CC
9337242_SilicoDArT	282767699	Allium cepa cultivar MRSPA ATP synthase subunit 6 (atp6) gene, complete cds; and unknown gene; mitochondrial	0.36	D2XT86	P92547	GO:0005739	Mitochondrion	CC
						GO:0005743	Mitochondrial inner membrane	CC
						GO:0015078	Hydrogen ion transmembrane transporter activity	MF
						GO:0015986	ATP synthesis coupled proton transport	BP
						GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
						GO:0045263	Proton-transporting ATP synthase complex, coupling factor F(o)	CC
9338024_SilicoDArT	970037987	PREDICTED: Solanum pennellii probable beta-1,3-galactosyltransferase 2 (LOC107023985), transcript variant X2, mRNA	0.01	K4CFD1	A8MRC7	GO:0000139	Golgi membrane	CC
						GO:0005794	Golgi apparatus	CC
						GO:0006486	protein glycosylation	BP
						GO:0008378	galactosyltransferase activity	MF
						GO:0016020	membrane	CC
						GO:0016021	integral component of membrane	CC
						GO:0016740	transferase activity	MF
						GO:0016757	transferase activity, transferring glycosyl groups	MF
9338065_SilicoDArT	1052171590	PREDICTED: Phoenix dactylifera L-arabinokinase-like (LOC103701454), mRNA	0.01	M0RKQ6	O23461	GO:0005524	ATP binding	MF
						GO:0005737	Cytoplasm	CC
						GO:0008152	Metabolic process	BP

						GO:0016301	Kinase activity	MF
						GO:0016310	Phosphorylation	BP
						GO:0016773	Phosphotransferase activity, alcohol group as acceptor	MF
9338202_SilicoDArT	955304399	PREDICTED: Glycine max 40S ribosomal protein S29 (LOC100775561), mRNA	0.00	A0A0A0L3R9	Q680P8	GO:0003735	Structural constituent of ribosome	MF
						GO:0005622	Intracellular	CC
						GO:0005840	Ribosome	CC
						GO:0006412	Translation	BP
9338212_SilicoDArT	469402936	Allium cepa AcRAD21-1 gene for cohesin subunit RAD21-1, partial cds, cultivar: Cheonjudaego	0.00	M5A7M8	Q8W1Y0	GO:0000228	Nuclear chromosome	CC
9338332_SilicoDArT	1083885722	PREDICTED: Nicotiana tomentosiformis ammonium transporter 3 member 1-like (LOC104107316), mRNA	0.01	K4CLT4	Q9M6N7	GO:0006810	Transport	BP
						GO:0008519	Ammonium transmembrane transporter activity	MF
						GO:0015696	Ammonium transport	BP
						GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
						GO:0072488	Ammonium transmembrane transport	BP
9339167_SilicoDArT	1052192815	PREDICTED: Phoenix dactylifera AP2-like ethylene-responsive transcription factor AIL5 (LOC103717110), mRNA	0.18	A9JQY7	Q6PQQ3	GO:0003677	DNA binding	MF
						GO:0003700	Transcription factor activity, sequence-specific DNA binding	MF
						GO:0005634	Nucleus	CC
						GO:0006351	Transcription, DNA-templated	BP
						GO:0006355	Regulation of transcription, DNA-templated	BP
						GO:0007275	Multicellular organism development	BP
9340282_SilicoDArT	1012205831	PREDICTED: Arachis duranensis sedoheptulose-1,7-bisphosphatase, chloroplastic (LOC107458166), mRNA	0.10	V7D2Q4	P46283	GO:0005975	Carbohydrate metabolic process	BP
						GO:0016311	Dephosphorylation	BP
						GO:0016791	Phosphatase activity	MF
						GO:0042578	Phosphoric ester hydrolase activity	MF

9340864_SilicoDArT	49781342	Allium cepa flavonol synthase gene, complete cds	0.01	Q6DTI1	Q96330	GO:0016491	Oxidoreductase activity	MF
						GO:0046872	Metal ion binding	MF
						GO:0055114	Oxidation-reduction process	BP
9341417_SilicoDArT	828303965	PREDICTED: Cicer arietinum glycogen synthase kinase-3 homolog MsK-1 (LOC101500512), transcript variant X3, mRNA	0.45	C7AE95	P43288	GO:0055114	Oxidation-reduction process	BP
						GO:0000166	Nucleotide binding	MF
						GO:0004672	Protein kinase activity	MF
						GO:0004674	Protein serine/threonine kinase activity	MF
						GO:0005524	ATP binding	MF
						GO:0006468	Protein phosphorylation	BP
						GO:0016301	Kinase activity	MF
						GO:0016310	Phosphorylation	BP
9341988_SilicoDArT	1130672164	PREDICTED: Elaeis guineensis glucose-6-phosphate 1-dehydrogenase, chloroplastic-like (LOC105051009), transcript variant X2, mRNA	0.19	A0A199W4P0	Q8L743	GO:0004345	Glucose-6-phosphate dehydrogenase activity	MF
						GO:0005975	Carbohydrate metabolic process	BP
						GO:0006006	Glucose metabolic process	BP
						GO:0016491	Oxidoreductase activity	MF
						GO:0050661	NADP binding	MF
						GO:0055114	Oxidation-reduction process	BP
9341999_SilicoDArT	166342	Allium sativum chitinase mRNA, 3' end	0.23	Q38776	P19171	GO:0004568	Chitinase activity	MF
						GO:0005975	Carbohydrate metabolic process	BP
						GO:0006032	Chitin catabolic process	BP
						GO:0008061	Chitin binding	MF
						GO:0016998	Cell wall macromolecule catabolic process	BP
9342092_SilicoDArT	747087535	PREDICTED: Sesamum indicum malate dehydrogenase (LOC105171782), mRNA	0.19	A0A067KZE2	P57106	GO:0003824	Catalytic activity	MF
						GO:0005975	Carbohydrate metabolic process	BP
						GO:0006099	Tricarboxylic acid cycle	BP

						GO:0006108	Malate metabolic process	BP
						GO:0016491	Oxidoreductase activity	MF
						GO:0016615	Malate dehydrogenase activity	MF
						GO:0016616	Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	MF
						GO:0019752	Carboxylic acid metabolic process	BP
						GO:0030060	L-malate dehydrogenase activity	MF
						GO:0055114	Oxidation-reduction process	BP
9342221_SilicoDArT	1050573319	PREDICTED: Gossypium arboreum sterol 3-beta-glucosyltransferase UGT80B1 (LOC108453732), transcript variant X4, mRNA	0.25	A0A067KZE2		GO:0005975	Carbohydrate metabolic process	BP
						GO:0008152	Metabolic process	BP
						GO:0016740	Transferase activity	MF
						GO:0016758	Transferase activity, transferring hexosyl groups	MF
						GO:0030259	Lipid glycosylation	BP
9343366_SilicoDArT	927028619	Allium sativum AsFMO1 mRNA for S-allyl-L-cysteine S-oxygenase, complete cds	0.37	A0A0M4U3V7	Q9FWW9	GO:0004497	Monooxygenase activity	MF
						GO:0004499	N, N-dimethylaniline monooxygenase activity	MF
						GO:0016491	Oxidoreductase activity	MF
						GO:0050660	Flavin adenine dinucleotide binding	MF
						GO:0050661	NADP binding	MF
						GO:0055114	Oxidation-reduction process	BP
9344094_SilicoDArT	828303334	PREDICTED: Cicer arietinum serine/threonine protein phosphatase 2A regulatory subunit B"beta-like (LOC101505006), mRNA	0.00	G7I4V6	Q5QIT3	GO:0005509	Calcium ion binding	MF
9344172_SilicoDArT	1081749611	Allium cepa fructan: fructan 6G-fructosyltransferase mRNA, complete cds	0.01	A0A125SXW6		GO:0004553	Hydrolase activity, hydrolyzing O-glycosyl compounds	MF
						GO:0005975	Carbohydrate metabolic process	BP
						GO:0008152	Metabolic process	BP
						GO:0016020	Membrane	CC

						GO:0016021	Integral component of membrane	CC
						GO:0016740	Transferase activity	MF
						GO:0016787	Hydrolase activity	MF
						GO:0016798	Hydrolase activity, acting on glycosyl bonds	MF
9344888_SilicoDArT	1040902836	PREDICTED: Daucus carota subsp. sativus F-box protein PP2-A15 (LOC108228058), mRNA	0.46	Q9LF92	Q9LF92	GO:0030246	Carbohydrate binding	MF
9345409_SilicoDArT	148872676	Allium cepa var. aggregatum lipid transfer protein 4 gene, complete cds	0.01	P27056		GO:0006810	Transport	BP
						GO:0006869	Lipid transport	BP
						GO:0008289	Lipid binding	MF
						GO:0016020	Membrane	CC
9345409_SilicoDArT	171221510	Allium cepa antimicrobial peptide mRNA, partial cds	0.01	Q41258		GO:0006952	Defense response	BP
						GO:0031640	Killing of cells of another organism	BP
						GO:0042742	Defense response to bacterium	BP
						GO:0050832	Defense response to fungus	BP
9345778_SilicoDArT	1083874822	PREDICTED: Nicotiana tomentosiformis DUF21 domain-containing protein At4g14240-like (LOC104101106), mRNA	0.01	A0A0V0IDD1	Q67XQ0	GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
9345846_SilicoDArT	1040916157	PREDICTED: Daucus carota subsp. sativus uncharacterized protein At1g04910-like (LOC108197265), mRNA	0.22	Q8W486		GO:0000139	Golgi membrane	CC
						GO:0005737	Cytoplasm	CC
						GO:0005768	Endosome	CC
						GO:0005794	Golgi apparatus	CC
						GO:0005802	Trans-Golgi network	CC
						GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
						GO:0016757	Transferase activity, transferring glycosyl groups	MF
9345909_SilicoDArT	1026013575	PREDICTED: Capsicum annuum phosphopantothenoylcysteine decarboxylase-like (LOC107864830), transcript variant X2, mRNA	0.27	K4CP90	Q9SWE5	GO:0003824	catalytic activity	MF
9346281_SilicoDArT	1002841322	PREDICTED: Oryza brachyantha thaumatin-like protein 1b (LOC102716100), mRNA	0.19	A0A199W7E4	Q9LNT0	GO:0016020	Membrane	CC

						GO:0016021	Integral component of membrane	CC
9346529_SilicoDArT	778662940	PREDICTED: Cucumis sativus beta-galactosidase-like (LOC101218515), mRNA	0.23	A0A0A0LWT7	Q9SCW1	GO:0004553	Hydrolase activity, hydrolyzing O-glycosyl compounds	MF
						GO:0005975	Carbohydrate metabolic process	BP
						GO:0008152	Metabolic process	BP
						GO:0016787	Hydrolase activity	MF
						GO:0016798	Hydrolase activity, acting on glycosyl bonds	MF
9346791_SilicoDArT	1109103029	PREDICTED: Camelina sativa probable pectinesterase/pectinesterase inhibitor 54 (LOC104770464), mRNA	0.25	A0A178UB30	Q3E989	GO:0004857	Enzyme inhibitor activity	MF
						GO:0005576	Extracellular region	CC
						GO:0005618	Cell wall	CC
						GO:0016787	Hydrolase activity	MF
						GO:0030599	Pectinesterase activity	MF
						GO:0042545	Cell wall modification	BP
						GO:0043086	Negative regulation of catalytic activity	BP
						GO:0045330	Aspartyl esterase activity	MF
						GO:0045490	Pectin catabolic process	BP
						GO:0071555	Cell wall organization	BP
9347750_SilicoDArT	1035395921	PREDICTED: Cucumis melo protein CHROMATIN REMODELING 4 (LOC103484261), transcript variant X4, mRNA	0.01	E5GCL1	F4KBP5	GO:0003677	DNA binding	MF
						GO:0005524	ATP binding	MF
						GO:0008270	Zinc ion binding	MF
						GO:0046872	Metal ion binding	MF
9348548_SilicoDArT	1130632797	PREDICTED: Elaeis guineensis chaperonin-like RbcX protein 2, chloroplastic (LOC105056619), mRNA	0.21	Q8L9X2	Q8L9X2	GO:0009507	Chloroplast	CC
						GO:0009536	Plastid	CC
						GO:0009570	Chloroplast stroma	CC
						GO:0044183	Protein binding involved in protein folding	MF
						GO:0061077	Chaperone-mediated protein folding	BP

9348634_SilicoDArT	1091648932	PREDICTED: Musa acuminata subsp. malaccensis calcium-transporting ATPase 1, endoplasmic reticulum-type-like (LOC104000928), mRNA	0.19	A0A199V502	Q9XES1	GO:0000166	Nucleotide binding	MF
						GO:0005524	ATP binding	MF
						GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
						GO:0016787	Hydrolase activity	MF
						GO:0046872	Metal ion binding	MF
9349862_SilicoDArT	793420965	Cucumis sativus probable sucrose-phosphate synthase 2 (LOC101208942), mRNA	0.01	S4TLQ4	Q94BT0	GO:0005985	Sucrose metabolic process	BP
						GO:0016157	Sucrose synthase activity	MF
						GO:0046524	Sucrose-phosphate synthase activity	MF
9349875_SilicoDArT	1109238907	PREDICTED: Ipomoea nil phosphoenolpyruvate carboxylase-like (LOC109188085), transcript variant X1, mRNA	0.01	A0A0V0IWX5	Q9MAH0	GO:0003824	Catalytic activity	MF
						GO:0006099	Tricarboxylic acid cycle	BP
						GO:0008964	Phosphoenolpyruvate carboxylase activity	MF
						GO:0015977	Carbon fixation	BP
9350048_SilicoDArT	1063488637	PREDICTED: Theobroma cacao polypyrimidine tract-binding protein homolog 1 (LOC18601541), mRNA	0.47	A0A061EEY4	Q9MAC5	GO:0000166	Nucleotide binding	MF
						GO:0003676	Nucleic acid binding	MF
9350049_SilicoDArT	1091726857	PREDICTED: Musa acuminata subsp. malaccensis polypyrimidine tract-binding protein homolog 1-like (LOC103987487), transcript variant X3, mRNA	0.49	A0A199V1H1		GO:0000166	Nucleotide binding	MF
						GO:0003676	Nucleic acid binding	MF
						GO:0009507	Chloroplast	CC
						GO:0009535	Chloroplast thylakoid membrane	CC
						GO:0009579	Thylakoid	CC
						GO:0015979	Photosynthesis	BP
						GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
9351005_SilicoDArT	923912599	PREDICTED: Brassica napus GDSL esterase/lipase LTL1-like (LOC106429781), mRNA	0.50	A0A078IIX2	Q9M8Y5	GO:0016788	Hydrolase activity, acting on ester bonds	MF

9351041_SilicoDArT	596187048	Prunus persica hypothetical protein (PRUPE_ppa005616mg) mRNA, complete cds	0.44	M5XSM0	Q8S9J6	GO:0004190	Aspartic-type endopeptidase activity	MF
						GO:0006508	Proteolysis	BP
						GO:0008233	Peptidase activity	MF
						GO:0016787	Hydrolase activity	MF
9351221_SilicoDArT	1108935616	PREDICTED: Beta vulgaris subsp. vulgaris probable protein S-acyltransferase 19 (LOC104895372), mRNA	0.46	A0A0J8CAT2	Q8L5Y5	GO:0008270	Zinc ion binding	MF
						GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
						GO:0016740	Transferase activity	MF
						GO:0016746	Transferase activity, transferring acyl groups	MF
						GO:0019706	Protein-cysteine S-palmitoyltransferase activity	MF
9351778_SilicoDArT	1025307851	PREDICTED: Nicotiana tabacum transcription factor MYB26-like (LOC107819474), transcript variant X2, mRNA	0.01	K4BMX3	Q9SPG3	GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	MF
						GO:0001135	Transcription factor activity, RNA polymerase II transcription factor recruiting	MF
						GO:0003677	DNA binding	MF
						GO:0005634	Nucleus	CC
						GO:0006357	Regulation of transcription from RNA polymerase II promoter	BP
						GO:0030154	Cell differentiation	BP
						GO:0043565	Sequence-specific DNA binding	MF
						GO:0044212	Transcription regulatory region DNA binding	MF
9352361_SilicoDArT	700253107	Asparagus officinalis MSH1 mRNA, partial cds	0.19	A0A097PJI7	Q84LK0	GO:0005524	ATP binding	MF
						GO:0006298	Mismatch repair	BP
						GO:0030983	Mismatched DNA binding	MF
9352365_SilicoDArT	1027091046	PREDICTED: Prunus mume E3 ubiquitin-protein ligase RHF2A (LOC103329838), mRNA	0.23	A0A0B2RPT0	Q9ZT42	GO:0008270	Zinc ion binding	MF
						GO:0016874	Ligase activity	MF

9352474_SilicoDArT	1091484874	PREDICTED: Eucalyptus grandis protein transport protein Sec61 subunit alpha (LOC104450102), mRNA	0.44	A0A059BV47		GO:0015031	Protein transport	BP
						GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
9353093_SilicoDArT	848861233	PREDICTED: Erythranthe guttatus 3-hydroxyisobutyryl-CoA hydrolase 1-like (LOC105952435), mRNA	0.49	A0A022RRA0	Q9LKJ1	GO:0003860	3-hydroxyisobutyryl-CoA hydrolase activity	MF
						GO:0016787	Hydrolase activity	MF
9353451_SilicoDArT	510122042	Allium cepa UFGT2 mRNA, complete cds	0.00	R9TPK0	Q9M156	GO:0008152	Metabolic process	BP
						GO:0016740	Transferase activity	MF
						GO:0016757	Transferase activity, transferring glycosyl groups	MF
						GO:0016758	Transferase activity, transferring hexosyl groups	MF
9353780_SilicoDArT	297809984	Arabidopsis lyrata subsp. lyrata protein translocase subunit secA chloroplast precursor, mRNA	0.21	D7M487	Q9SYI0	GO:0000166	Nucleotide binding	MF
						GO:0005524	ATP binding	MF
						GO:0005622	Intracellular	CC
						GO:0006605	Protein targeting	BP
						GO:0006810	Transport	BP
						GO:0006886	Intracellular protein transport	BP
						GO:0015031	Protein transport	BP
						GO:0016020	Membrane	CC
						GO:0017038	Protein import	BP
9353802_SilicoDArT	1052178563	PREDICTED: Phoenix dactylifera zinc finger protein CONSTANS-like (LOC103706571), transcript variant X4, mRNA	0.46	M0SH14		GO:0005622	Intracellular	CC
						GO:0008270	Zinc ion binding	MF
9354458_SilicoDArT	1098721498	PREDICTED: Juglans regia casein kinase 1-like protein 1 (LOC109009397), mRNA	0.25	W9QK02	Q9CAI5	GO:0000166	Nucleotide binding	MF
						GO:0004672	Protein kinase activity	MF
						GO:0004674	Protein serine/threonine kinase activity	MF
						GO:0005524	ATP binding	MF
						GO:0006468	Protein phosphorylation	BP
						GO:0016301	Kinase activity	MF

						GO:0016310	Phosphorylation	BP
9354508_SilicoDArT	747066802	PREDICTED: Sesamum indicum cellulose synthase A catalytic subunit 4 [UDP-forming] (LOC105163451), mRNA	0.08	A0A022S0V4	Q84JA6	GO:0005886	Plasma membrane	CC
						GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
						GO:0016740	Transferase activity	MF
						GO:0016757	Transferase activity, transferring glycosyl groups	MF
						GO:0016760	Cellulose synthase (UDP-forming) activity	MF
						GO:0030244	Cellulose biosynthetic process	BP
						GO:0046872	Metal ion binding	MF
						GO:0071555	Cell wall organization	BP
9354557_SilicoDArT	1130635044	PREDICTED: Elaeis guineensis BTB/POZ domain-containing protein At5g03250-like (LOC105032513), mRNA	0.20	Q9LYW0		GO:0016567	Protein ubiquitination	BP
9354568_SilicoDArT	42565431	Hyacinthus orientalis 40S ribosomal protein S23 mRNA, partial cds	0.33	Q5YJP7	Q9SF35	GO:0003735	Structural constituent of ribosome	MF
						GO:0005622	Intracellular	CC
						GO:0005840	Ribosome	CC
						GO:0006412	Translation	BP
9354630_SilicoDArT	242041800	Sorghum bicolor hypothetical protein, mRNA	0.21	C5WTM9	Q9LF33	GO:0003979	UDP-glucose 6-dehydrogenase activity	MF
						GO:0016491	Oxidoreductase activity	MF
						GO:0016616	Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	MF
						GO:0051287	NAD binding	MF
						GO:0055114	Oxidation-reduction process	BP
9354789_SilicoDArT	1098803480	PREDICTED: Juglans regia mitogen-activated protein kinase 8-like (LOC108980971), mRNA	0.23	M5WE81	Q9LV37	GO:0000165	MAPK cascade	BP
						GO:0000166	Nucleotide binding	MF
						GO:0004672	Protein kinase activity	MF
						GO:0004674	Protein serine/threonine kinase activity	MF

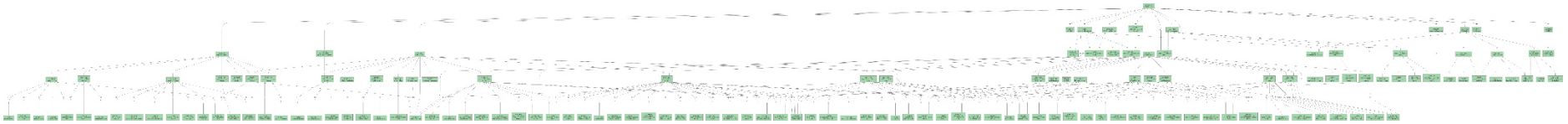
						GO:0004707	MAP kinase activity	MF
						GO:0005524	ATP binding	MF
						GO:0005622	Intracellular	CC
						GO:0006468	Protein phosphorylation	BP
						GO:0016301	Kinase activity	MF
						GO:0016310	Phosphorylation	BP
						GO:0016740	Transferase activity	MF
9354847_SilicoDArT	1098820251	PREDICTED: Juglans regia RNA polymerase II-associated protein 3 (LOC108987762), transcript variant X3, mRNA	0.18	A0A0B2QTW7		GO:0004721	Phosphoprotein phosphatase activity	MF
						GO:0006470	Protein dephosphorylation	BP
						GO:0016787	Hydrolase activity	MF
9354897_SilicoDArT	960458686	PREDICTED: Brachypodium distachyon ribulose-phosphate 3-epimerase, chloroplastic (LOC100826802), transcript variant X2, mRNA	0.01	M0YKB8	Q9SAU2	GO:0003824	Catalytic activity	MF
						GO:0004750	Ribulose-phosphate 3-epimerase activity	MF
						GO:0005975	Carbohydrate metabolic process	BP
						GO:0006098	Pentose-phosphate shunt	BP
						GO:0008152	Metabolic process	BP
						GO:0016853	Isomerase activity	MF
						GO:0016857	Racemase and epimerase activity, acting on carbohydrates and derivatives	MF
						GO:0046872	Metal ion binding	MF
9355046_SilicoDArT	1130627979	PREDICTED: Elaeis guineensis galacturonosyltransferase 8-like (LOC105035085), mRNA	0.01	M0T698	Q9LSG3	GO:0000139	Golgi membrane	CC
						GO:0005794	Golgi apparatus	CC
						GO:0016740	Transferase activity	MF
						GO:0016757	Transferase activity, transferring glycosyl groups	MF
						GO:0045489	Pectin biosynthetic process	BP
						GO:0047262	Polygalacturonate 4-alpha-galacturonosyltransferase activity	MF

						GO:0071555	Cell wall organization	BP
9355352_SilicoDArT	358248409	Glycine max 40S ribosomal protein S3-3-like (LOC100808705), mRNA	0.19	A0A0B2NWG8	Q9FJA6	GO:0003723	RNA binding	MF
						GO:0003735	Structural constituent of ribosome	MF
						GO:0005840	Ribosome	CC
						GO:0006412	Translation	BP
						GO:0015935	Small ribosomal subunit	CC
						GO:0030529	Intracellular ribonucleoprotein complex	CC
9355596_SilicoDArT	1105486800	PREDICTED: Vitis vinifera beta-galactosidase 3 (LOC100232848), mRNA	0.19	D7SP52	Q9SCV9	GO:0004553	Hydrolase activity, hydrolyzing O-glycosyl compounds	MF
						GO:0004565	Beta-galactosidase activity	MF
						GO:0005618	Cell wall	CC
						GO:0005773	Vacuole	CC
						GO:0005975	Carbohydrate metabolic process	BP
						GO:0008152	Metabolic process	BP
						GO:0009505	Plant-type cell wall	CC
						GO:0016787	Hydrolase activity	MF
						GO:0016798	Hydrolase activity, acting on glycosyl bonds	MF
						GO:0030246	Carbohydrate binding	MF
9355597_SilicoDArT	922534331	PREDICTED: Brassica oleracea var. oleracea beta-galactosidase 3-like (LOC106306365), mRNA	0.49	A0A0D3DHZ3		GO:0004553	Hydrolase activity, hydrolyzing O-glycosyl compounds	MF
						GO:0004565	Beta-galactosidase activity	MF
						GO:0005975	Carbohydrate metabolic process	BP
						GO:0008152	Metabolic process	BP
						GO:0016787	Hydrolase activity	MF
						GO:0016798	Hydrolase activity, acting on glycosyl bonds	MF
						GO:0030246	Carbohydrate binding	MF

9356271_SilicoDArT	1021570407	PREDICTED: Arachis ipaensis peroxidase A2-like (LOC107618757), mRNA	0.36	A0A0B2PDR7	Q42578	GO:0004601	Peroxidase activity	MF
						GO:0005576	Extracellular region	CC
						GO:0006979	Response to oxidative stress	BP
						GO:0016491	Oxidoreductase activity	MF
						GO:0020037	Heme binding	MF
						GO:0042744	Hydrogen peroxide catabolic process	BP
						GO:0046872	Metal ion binding	MF
						GO:0055114	Oxidation-reduction process	BP
						GO:0098869	Cellular oxidant detoxification	BP
9356301_SilicoDArT	769794621	PREDICTED: Amborella trichopoda BTB/POZ and TAZ domain-containing protein 3 (LOC18448701), transcript variant X3, mRNA	0.01	U5D3V0	Q9SYL0	GO:0003712	Transcription cofactor activity	MF
						GO:0004402	Histone acetyltransferase activity	MF
						GO:0005516	Calmodulin binding	MF
						GO:0005634	Nucleus	CC
						GO:0006355	Regulation of transcription, DNA-templated	BP
						GO:0008270	Zinc ion binding	MF
						GO:0009409	Response to cold	BP
						GO:0009553	Embryo sac development	BP
						GO:0009555	Pollen development	BP
						GO:0009651	Response to salt stress	BP
						GO:0009723	Response to ethylene	BP
						GO:0009751	Response to salicylic acid	BP
						GO:0016573	Histone acetylation	BP
						GO:0042542	Response to hydrogen peroxide	BP
9356799_SilicoDArT	296184580	Dimocarpus longan cultivar Honghezi 14-3-3 family protein gene, complete cds	0.40	D4P505	P42644	GO:0019904	Protein domain specific binding	MF
9356800_SilicoDArT	1040831491	PREDICTED: Daucus carota subsp. sativus 14-3-3-like protein (LOC108204889), mRNA	0.33	A0A166I732		GO:0019904	Protein domain specific binding	MF
9357227_SilicoDArT	1000986181	PREDICTED: Ricinus communis peroxidase 64 (LOC8267824), mRNA	0.49	B9R8E4	Q43872	GO:0004601	Peroxidase activity	MF

						GO:0005576	Extracellular region	CC
						GO:0006979	Response to oxidative stress	BP
						GO:0016491	Oxidoreductase activity	MF
						GO:0020037	Heme binding	MF
						GO:0042744	Hydrogen peroxide catabolic process	BP
						GO:0046872	Metal ion binding	MF
						GO:0055114	Oxidation-reduction process	BP
						GO:0098869	Cellular oxidant detoxification	BP
9357457_SilicoDArT	922467943	PREDICTED: Brassica oleracea var. oleracea protein transport protein Sec61 subunit beta-like (LOC106339764), mRNA	0.00	A0A0D3BQ04	P38389	GO:0005784	Sec61 translocon complex	CC
						GO:0006886	Intracellular protein transport	BP
						GO:0016020	Membrane	CC
						GO:0016021	Integral component of membrane	CC
9357544_SilicoDArT	778682834	PREDICTED: Cucumis sativus clustered mitochondria protein (LOC101207687), transcript variant X1, mRNA	0.01	A0A0A0L9W0	F4J5S1	GO:0003723	RNA binding	MF
						GO:0005737	Cytoplasm	CC
						GO:0007005	Mitochondrion organization	BP
						GO:0048312	Intracellular distribution of mitochondria	BP
9358018_SilicoDArT	1050591662	PREDICTED: Gossypium arboreum protein mago nashi homolog (LOC108461752), mRNA	0.23	A0A0D2R0E2	O23676	GO:0005634	Nucleus	CC
9358018_SilicoDArT	1063871953	Phialophora attae hypothetical protein mRNA	0.23	A0A0D2R0E2		GO:0005634	Nucleus	CC
9358259_SilicoDArT	703088396	Morus notabilis Germination-specific cysteine protease 1 partial mRNA	0.30	W9QU40	Q94B08	GO:0006508	Proteolysis	BP
						GO:0008233	Peptidase activity	MF
						GO:0008234	Cysteine-type peptidase activity	MF
						GO:0016787	Hydrolase activity	MF
9358310_SilicoDArT	1052183060	PREDICTED: Phoenix dactylifera dehydrogenase/reductase SDR family member 12 (LOC103709761), mRNA	0.18	NA		NA	NA	NA
9358363_SilicoDArT	836002706	PREDICTED: Setaria italica 26S proteasome non-ATPase regulatory subunit 8 homolog A-like (LOC101769553), mRNA	0.32	W5BH47	Q9SGW3	GO:0005838	Proteasome regulatory particle	CC
						GO:0006508	Proteolysis	BP

9359097_SilicoDArT	1111112471	PREDICTED: Nicotiana attenuata dnaJ protein ERDJ3B-like (LOC109227226), mRNA	0.45	K4AXK6	Q9LZK5	GO:0006457	Protein folding	BP
						GO:0051082	Unfolded protein binding	MF
9360464_SilicoDArT	741985377	Allium cepa dihydroflavonol 4-reductase (DFR-A) gene, DFR-ADTP allele, complete cds; and transposon AcCACTA1, complete sequence	0.26	X5CJ79	P51102	GO:0003824	Catalytic activity	MF
						GO:0050662	Coenzyme binding	MF
9360611_SilicoDArT	662247390	Nicotiana tabacum HSP18.9 mRNA, complete cds	0.05	P19037	P19037	GO:0005515	Protein binding	MF
						GO:0005737	Cytoplasm	CC
						GO:0009408	Response to heat	BP
						GO:0009644	Response to high light intensity	BP
						GO:0010286	Heat acclimation	BP
						GO:0042542	Response to hydrogen peroxide	BP
9360611_SilicoDArT	99033682	Agave tequilana clone 1.8 chaperone mRNA, complete cds	0.05	Q84Q72		GO:0005737	Cytoplasm	CC
						GO:0009408	Response to heat	BP
						GO:0042542	Response to hydrogen peroxide	BP
						GO:0045471	Response to ethanol	BP
						GO:0046685	Response to arsenic-containing substance	BP
						GO:0046686	Response to cadmium ion	BP
						GO:0046688	Response to copper ion	BP
9360611_SilicoDArT	326528088	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone: NIASHv1064A22	0.05	P19036	P19036	GO:0005515	Protein binding	MF
						GO:0005737	Cytoplasm	CC
						GO:0009408	Response to heat	BP
9360612_SilicoDArT	99033708	Agave tequilana clone 7.8 chaperone mRNA, complete cds	0.05	P27880		GO:0005737	Cytoplasm	CC
9360612_SilicoDArT	326504765	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone: NIASHv3095H06	0.05	Q84Q77	A0A178V6V7	GO:0005634	Nucleus	CC
						GO:0009408	Response to heat	BP
9360612_SilicoDArT	1109265466	PREDICTED: Ipomoea nil 18.2 kDa class I heat shock protein-like (LOC109153464), mRNA	0.05	P27880		GO:0005737	Cytoplasm	CC



Supplementary Material 2.3. Blast2GO graph, including all BP terms found after GO analyses. BP are among them by “is a”, “part of”, “regulates” or “positively regulates”. From left to right: processes related to Response to stimulus, Cellular processes, Regulations. Other signaling. Developmental processes, Single or multicellular processes, Metabolic processes, Cellular localization, and Cell killing processes.This figure can be seen with higher resolution in the CD attached.

CHAPTER 3. Genotyping Worldwide Olive Germplasm Bank varieties by High-Resolution Melting (HRM)

3.1. Abstract

Olive tree (*Olea europaea* L.) is a Mediterranean species which is currently grown also in America, Asia, and Oceania. Belonging to *Oleaceae* family, with more than 30 genera and 600 species, *Olea* genus has more than 30 species. *O. europaea* is a perennial diploid (2n = 46) species with 1.38 Gbp haploid genome, which is highly heterozygous. This is enhanced by its predominantly allogamous reproduction. The relevance of olive tree lies in the fact that it is one of the most important oil crops. Oil bioactivity, and its well-balanced fatty-acid composition, provides olive oil with many interesting properties such as cardiovascular protection, being antioxidant, preventing cancer, and anti-inflammatory activities. This has caused an increment of olive oil consumption, even in non-traditional producer and consumer countries. In Spain, the Worldwide Olive Germplasm Bank of Córdoba (WOGBC) has more than 400 olive-tree varieties. In general, germplasm banks constitute a reservoir of genetic diversity and variability for genetic breeding. In the case of olive-tree varieties, several molecular-marker approaches have already been performed. Among others are RFLP, RAPD, SCAR, SSR, ISSR, AFLP, SNP, EST or DArT. Moreover, transcriptomes of some olive cultivars like Picual, Arbequina, and Lechín de Sevilla have been recently sequenced. In this chapter, HRM analyses have been performed in order to genetically assess the WOGBC. HRM technique has the advantage of being a fast and simple, "closed-tube" approach, with low sample and reagent requirements, being also highly sensitive and specific. Five EST and one Sequence Tagged Site (STS) marker were used in order to perform Unweighted Pair-Group Method with Arithmetic Mean (UPGMA) analysis and evaluate genetic diversity of 83 samples of the WOGBC. As a result, most samples were grouped in the dendrogram according to their geographical origin. In general, results were in agreement with previous works with WOGBC samples. HRM proved to be an effective tool to perform genetic-diversity analysis, in a fast and affordable way, in a species with available genetic information. To the best of our knowledge, this is the first HRM genotyping of olive tree, applied to a large number of WOGBC accessions from many different countries. This methodology can be useful to identify agronomical traits of interest, including production and resistance to abiotic and biotic stresses.

3.2. Introduction

3.2.1. Olive tree: botanic, taxonomic and health aspects

Olive tree (*Olea europaea* subsp. *europaea* var. *europaea*; usually known as *Olea europaea*) originated in the Mediterranean Basin, being also currently grown in America, Asia, and Oceania. It belongs to *Oleacea*e family, comprising more than 30 genera and 600 species (Cronquist, 1981). *Olea* genus has more than 30 species found in Africa, Asia, Europe, and Oceania. Olive tree is a perennial (Green, 2002) diploid (2n = 46) species, with a relatively small genome of 1.38 Gbp (haploid) that is highly heterozygous. This is enhanced by the predominant allogamy of such species (Cruz et al., 2016).

Olive tree (the only cultivated species of its genus) is a relevant agri-food plants, being one of the most important oil crops. Indeed, olive oil is a fruit juice that has many health benefits. This includes cardiovascular protection, being antioxidant, preventing cancer, and anti-inflammatory activity. Olive-oil bioactivity is due to secondary metabolites, including phenolic compounds such as lignans and secoiridoids. It also has a well-balanced fatty-acid composition, being rich in monounsaturated oleic acid. This makes it particularly resistant to high-temperature cooking (being therefore well suited for deep frying) and oxidation (Beauchamp et al., 2005; Pérez-Jiménez et al., 2007; Domínguez-García et al., 2012; Servili et al., 2013). These beneficial properties have increased worldwide olive-oil consumption, even in non-traditional producing or consuming countries, like Australia, Japan, and United States of America (USA) (Bracci et al., 2011; Servili et al., 2013). In fact, the Food and Drug Administration (FDA) of USA <http://www.fda.gov> has granted olive oil a qualified health-claim label, for protecting against coronary-heart disease (Docket No. 2003Q-0559). Besides its healthy properties as food, olive-tree derived products (like the ones obtained from leaves and as by-products of olive oil production) are also being exploited by pharmaceutical and cosmetic industries, mainly for their antioxidant and wound-healing properties (Rodrigues et al., 2015).

3.2.2. Characterization of olive-tree varieties

The vast majority of olive-tree varieties have been generated by traditional breeding, for instance, using clonal selection, cross breeding or mutagenesis, and are bounded to their original geographic area (Besnard et al., 2001; Fabbri et al., 2009). Nevertheless, ought to multiple factors such as olive-tree reproductive system (vegetatively propagated), longevity, phenotypic plasticity, and climatic adaptation, there is a great diversity among cultivars. This may generate synonymy and homonymy cases (Mariotti et al., 2016). Olive-tree cultivars originally came from Southern Europe, where 70% of worldwide olive-oil is produced and consumed. In Spain, olive tree is the woody crop with the highest cultivated area (approximately, 2.5 million hectares), which represents 5% of total cultivated area. Specifically, in Andalusia, this area is 1.5 million hectares, representing 18% of total cultivated surface in the country (Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente, 2015).

The Worldwide Olive Germplasm Bank of Córdoba (WOGBC) in Spain was founded in 1970. It has more than 400 varieties, being the first and largest olive-tree collection (Caballero et al., 2005). Worldwide, in total there are more than 1,200 cultivated varieties, of which, approximately 540 commercial olive-tree varieties were from Italy, 180 from Spain, 90 from France, and 50 from Greece (Baldoni and Belaj, 2009). Germplasm banks may constitute invaluable reservoirs of genetic diversity and variability for genetic breeding. Yet, classification criteria have been traditionally carried out using morphologic or agronomic data. This may increase the probability of storing redundant samples and of having cases of synonymy. Thus, it is vital to genetically assess germplasm banks, to create core collections and to facilitate breeding programs (Zhao et al., 2010). Studies on agronomical traits and genetic diversity allow to identify biodiversity, further optimizing germplasm management and conservation (Nybom and Bartish, 2000; Mohammadi and Prasanna, 2003; Simko et al., 2012).

Different molecular-marker approaches can be used to assess genetic variability of populations, including olive tree. Among others, they include proteic/enzymatic polymorphisms, AFLP (Angiolillo et al., 1999), RFLP (Besnard et al., 2001), RAPD (Besnard et al., 2001), Sequence-Characterized Amplified Regions (SCAR) (Busconi et

al., 2006), Inter Simple-Sequence Repeats (ISSR) (Gomes et al., 2008), SNP (Muleo et al., 2009), SSR (Belaj et al., 2012), DArT (Belaj et al., 2012), and EST markers (Mariotti et al., 2016). EST markers are short complementary DNA (cDNA) sequences, randomly generated from cDNA clones. They represent a fast and efficient method for gene profiling, gene mapping, and genotyping. This is particularly relevant when there is no or scarce sequence information. Furthermore, they avoid problems associated with genome size or transposable-element repetitions, and its use allows a rapid and cost-effective genotyping. Notwithstanding, some caveats should be taken into account. Firstly, as they come from transcribed DNA (that is, expressed genes), they may not contain a broad representation of the full genome content. Secondly, due to the origin of this technology (random sequencing of cDNA fragments), sequence accuracy may be reduced, as some errors could occur during cDNA synthesis or sequencing (Dorado et al., 2015). This could be solved by resequencing EST databases. Conversely, that is time consuming not being always possible. Finally, a third source of error may stem from human manipulation (data processing). Nevertheless, despite these potential drawbacks, EST databases have proved to be a valuable resource of molecular markers for several genetic purposes. Furthermore, not only EST markers are useful when there is no genetic information about the species, but they also have helped to increase our knowledge of interesting agronomic traits for breeding programs (Alba et al., 2004).

In addition, transcriptomes of some olive-tree varieties have also been assembled and annotated. They include Picual, Arbequina, Lechín de Sevilla cultivars, and seedlings from segregating progeny of Picual × Arbequina cross (Muñoz-Mérida et al., 2013). Additionally, the first draft of olive-tree genome (Farga variety) has been recently published (Cruz et al., 2016). These works represent tools for a better genetic characterization of olive-tree cultivars, including the search for interesting agronomical traits. Among them are the ones related to flowering and fruit production. Also relevant are the ones related to resistance to abiotic (temperature, drought, etc.) and biotic stresses (diseases and pests). Indeed, breeding to improve environmental adaptation is particularly relevant in the current trend of climate change and global warming.

3.2.3. High-Resolution Melting analyses

HRM allows to analyze thermal-melting profiles of previously amplified DNA fragments. It is a versatile technique applicable to different kinds of molecular markers; for instance, SSR, EST or SNP. One of its advantages is that it is a "closed-tube" approach. Hence, gel or capillary electrophoresis are not necessary, reducing time and cost during analysis processes. Most importantly, that avoids cross contaminations after DNA amplifications. Additionally, genotyping can be semi-automatically accomplished by the same software used for quantitative Real-Time PCR (qRT-PCR). In short, PCR is performed in presence of a fluorescent intercalating dyes that specifically bind to double-stranded DNA (dsDNA). They emit a strong fluorescence in comparison to their low fluorescence when unbound. Therefore, fluorescence levels allow to monitor the PCR-amplification process. When PCR has finished, amplicons are subjected to increased temperature steps that gradually denature DNA. While dsDNA is denaturing, the emitted fluorescence weakens. By plotting fluorescence against temperature, melting curves are generated and visualized in real time. Their shapes may be unique, depending on several factors like amplicon length, amplified sequence and CG content versus AT one. Thus, it is fast and simple, with low sample and reagent requirements, being highly sensitive and specific. Indeed, it allows to discriminate small sequence differences, like SNP or single-base insertions/deletions (indels). Nevertheless, HRM profiles may be similar in some instances, hindering genotype differentiation. In any case, HRM approach is usually highly recommended to study biodiversity by molecular markers (Distefano et al., 2013).

HRM analyses have been extensively used in plants. This includes fingerprinting studies with SNP and SSR in many genus, such as *Arabidopsis, Capsicum*, *Citrus*, *Citrullus, Cucurbita, Ficus, Jatropha, Origanum, Oryza, Prunus, Rubus, Salvia, Solanum, Sorghum,* or *Vitis,* among others (Distefano et al., 2013, 2015; Caruso et al., 2014; Simko, 2016). In addition, HRM analyses have been used to develop trait-linked markers and to map genes (Bracci et al., 2011; Bushakra et al., 2012). Yet, few HRM analyses have been done to genotype and evaluate genetic diversity in olive-tree varieties (Mackay et al., 2008; Las Casas et al., 2014). Most diversity studies are restricted to specific locations, such as Italy and Greece (Muleo et al., 2009; Xanthopoulou et al., 2014). Furthermore, HRM technology has been used for olive-oil traceability. This allows to identify varieties, certify authenticity, and detect putative adulterations with oils from other species (Vietina et al., 2013; Montemurro et al., 2015; Pasqualone et al., 2015).

Recently, some works in olive tree have combined molecular characterization by HRM with chemical analyses (Pasqualone et al., 2016).

3.2.4. Objective

The aim of this chapter is to describe a cost-effective and "closed-tube" method of genotyping olive-tree varieties by HRM.

3.3. Materials and methods

3.3.1. Plant material and DNA isolation

A total of 83 olive-tree cultivars from the WOGBC were used in the analyses (Table 3.1). DNA was isolated from leaves by Cetyl Trimethyl Ammonium Bromide (CTAB) protocol (Murray and Thompson, 1980) as further optimized (Hernandez et al., 2001). Samples were dissolved in Tris [tris(hydroxymethyl)aminomethane] - ethylene diamine-tetraacetic acid (EDTA; TE buffer with pH 8) and stored at –20 °C. Isolated DNA was quantified by NanoDrop 2000c from Thermo Fisher Scientific (Waltham, MA, USA). Samples were subjected to gel electrophoresis using 0.8% (w/v) agarose from Sigma-Aldrich (San Luis, MO, USA). DNA was stained with GelGreen from Biotium (Hayward, CA, USA). Bands were visualized under blue light using a DR195M "Dark Reader" transilluminator from Clare Chemical Research (Dolores, CO, USA).

Table 3.1. Cultivation area and number of analyzed samples.

Cultivation area	Number of samples
Algeria	1
Chile	1
Croatia	4
Egypt	1
France	4
Greece	2
Israel	2
Italy	6
Lebanon	1
Morocco	3
Portugal	1
Spain	42
Syria	4
Tunisia	1
Turkey	6

3.3.2. Genotyping by HRM analyses

Five primer pairs previously developed in the research group using EST (M1 to M5) sequences and one STS sequence (M6) were used for HRM genomic profiling. PCR were performed with Type-it HRM PCR Kit from Qiagen (Hilden, Germany). Briefly,

PCR reactions contained 10 µl of 2X Master Mix (HotStarTaq Plus DNA Polymerase, Type-it HRM PCR Buffer with EvaGreen dye, Q-Solution and dNTP mix), 0.7 µM of each primer, 40 ng of genomic DNA (gDNA) and diethyl pyrocarbonate (DEPC)-treated water, up to a final reaction volume of 20 µl. PCR was performed in Rotor-Gene 6000 thermal cycler from Qiagen. PCR amplification profile included: i) initial DNA-polymerase activation at 95 °C for 5 min; ii) 48 cycles with denaturation at 95 °C for 20 s, annealing at 72 °C for 30 s and extension at 72 °C during 15 s; iii) 95 °C for 1 min to denature DNA; and iv) 40 °C for 1 min to renature DNA. HRM analyses were performed using a ramp from 55 °C to 99 °C, with acquisitions for each 0.1 °C temperature increment every 2 s. No template control was included for each run.

Genotyping was performed with Rotor-Gene software version 1.7. Percentage of normalized fluorescence signal was plotted against temperature to show HRM curves. Then, one sample was assigned as "reference genotype" from each curve pattern, and samples were automatically assigned to genotype groups with similar melting curves. Chosen percentage of confidence to accept curve pairing was 90%. Finally, a matrix table with all samples and markers was created for further analyses. Missing rates were lower than 5% for every marker. All primers were checked to be polymorphic before starting genetics analyses.

3.3.3. Genetic diversity analysis

UPGMA analysis were performed, in order to study genetic diversity. A dendrogram was generated with all samples of different countries. A Nei distance matrix (Nei et al., 1983) was calculated with PowerMarker software version 3.25 (Liu and Muse, 2005) to create the tree. Then, the matrix was used as input in Phylip version 3.696 from NEIGHBOUR package (Felsenstein, 1989). Statistical significance of dendrograms was tested with Cophenetic Correlation-Coefficient (CCC), using a Visual Basic Macro in Excel (Dighe et al., 2004). Final version of tree was edited with MEGA software version 7 (Kumar et al., 2016).

3.4. Results

3.4.1. HRM genotyping

A total of 83 WOGBC samples were genotyped by HRM to assess genetic diversity and structure. They were subjected to PCR amplification and their normalized HRM curves were analyzed. HRM profiles are shown in Figure 3.1 and genotyping results are shown in Table 3.2 As it can be seen, M2 marker showed the greatest number of alleles (eight) followed by M3 (six) and M1 (four). The other markers had three alleles. Mean Polymorphic-Information Content (PIC) was 0.57, with values ranging from 0.45 to 0.80. Mean genetic diversity was 0.63, with a range of 0.52 to 0.82 (Table 3.2).

Table 3.2. Allele comparison and genotypic information for plastid EST (M1-M5) and STS (M6) markers.

Marker	Individuals analyzed	Number of alleles	Genetic diversity	PIC
M1	83	4	0.63	0.56
M2	83	8	0.82	0.80
M3	83	6	0.66	0.63
M4	83	3	0.52	0.45
M5	83	3	0.52	0.45
M6	83	3	0.63	0.56
Mean	83	4.5	0.63	0.57

3.4.2. Genetic diversity assessment

UPGMA dendrogram analyses divided samples in three main clusters (Fig. 3.2). Clusters 1 to 3 were highlighted in green, blue, and red, respectively. Cluster 2 was subdivided into two subclusters (differentiated in dark and light blue). CCC for UPGMA dendrogram was 0.81. Interestingly, some samples were grouped together according to their geographical origin. For instance, POR01 and POR02, both coming from Portugal were grouped in dark-blue cluster. French varieties were mainly grouped in blue cluster. Turkish (TUR01 to TRU04 and TUR06), and Syrian (SYR01 to SYR03) varieties were mostly located in light-blue cluster. Croatian cultivars were in red cluster, except for CRO01, which was clustered in another group (blue). Greek samples were also found in red cluster. ARG01 and MOR03 were also found together, being indeed considered synonymous accessions (Trujillo et al., 2014). Another case of synonymy accessions

described was SPA13 and SPA27 (Trujillo et al., 2014), being likewise clustered together in the dendrogram.

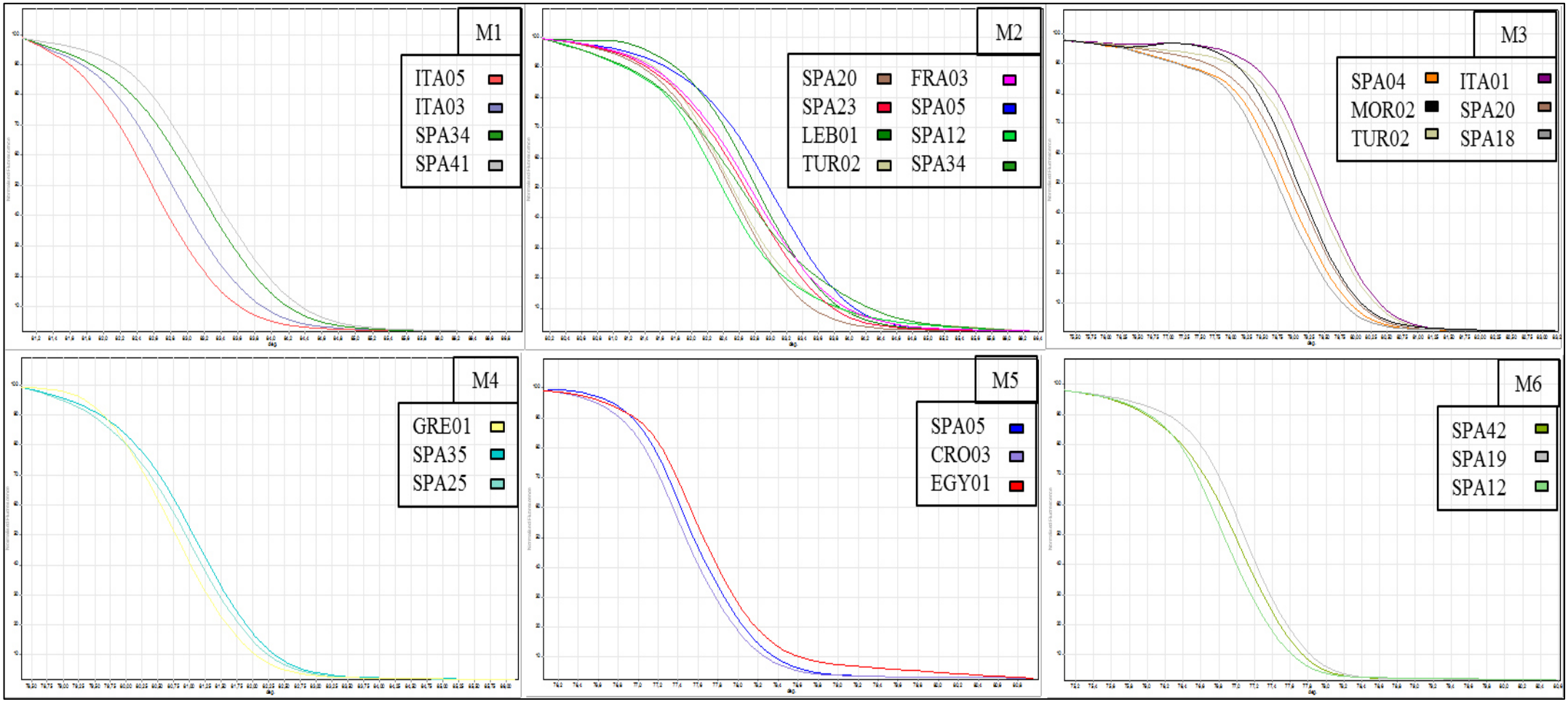


Figure 3.1. HRM plots for samples selected as model curves for each marker. Curve's color (described in each graph) corresponds to the chosen accessions in each run. Samples' codes are described in Supplementary Material 3.1.

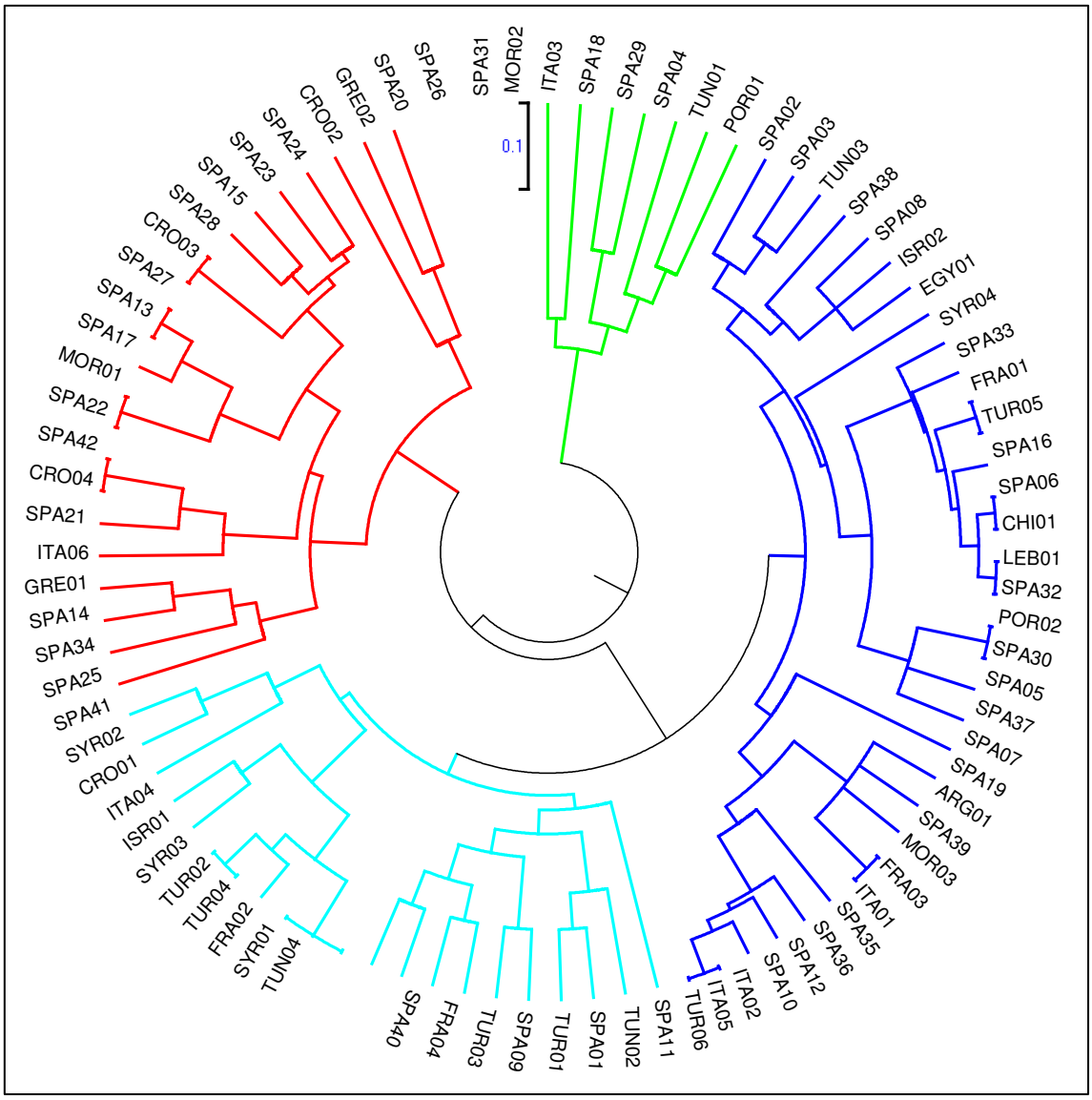


Figure 3.2. UPGMA dendrogram. The scale indicates branch length. Cultivar names are indicated with short version, complete names can be found in Table 3.1. CCC value was 0.81. Clusters 1 to 3 are highlighted in green, blue (dark blue and light blue for each subcluster) and red, respectively.

Conversely, not all samples were grouped according to their putative origin. For instance, Israeli, French, and Tunisian samples were found in all clusters. Spanish and Italian cultivars which were the most abundant, were also spread through all clusters.

3.5. Discussion

In this chapter, HRM analysis has been used as a "closed-tube" technique to show its utility for olive genotyping. In this case, six molecular markers have been used to genotype 83 WOGBC samples. Although a limited number of samples were analyzed using only six markers, some relationships among samples could be found. In any case, the main objective of this chapter was not to study genetic diversity, but to show the usefulness of HRM genotyping as a cost- and time-effective approach.

Interestingly, in general, samples grouped in accordance to their geographical origin in the UPGMA analyses. This was the case for varieties from Portugal, Turkey, Syria, and Croatia. In addition, synonyms varieties (Trujillo et al., 2014) were also clustered together. For instance, MOR03 and ARG01, or SPA13 and SPA27. Notwithstanding, some samples were not clustered according to such criterion. For example, samples from Spain, Italy, Israel, France, and Tunisia were found across the different dendrogram clusters. This is not surprising and it could be due to several reasons. Firstly, the low number of analyzed markers (6). Secondly, the number of Italian –and mostly Spanish– samples was higher in comparison to other countries. Hence, that may increase the probability of finding them in different clusters. Additionally, an accession collected at a specific country may have been originated in other, albeit such fact may be unknown. This could be due to exchanges carried out since ancient times among farmers, in order to select olive-tree varieties with desired characteristics. These exchanges were probably more frequent between nearby areas. This is due to convenience, but also because local varieties are usually better adapted to their cultivation areas (Bartolini et al., 2002). Hence, this could hinder the possibility of differentiating samples of nearby geographic-regions, or with similar climatic conditions. Therefore, it is possible to find accessions across different clusters from countries with low representation of varieties, as in the case of Israel, France or Tunisia. In any case, these issues highlight the relevance of molecular markers, to properly genotype and identify accessions in germplasm banks.

Despite the above limitations, the genetic diversity and structure results obtained from one STS and five EST markers were interesting. UPGMA consistency was supported by CCC showing a dendrogram value of 0.81, which is a good fit. Therefore, dendrogram accurately preserved pairwise distances between original data. Generated clusters, as expected, had similarities and differences with previous studies on genetic diversity by molecular markers, as reported for WOGBC by other authors (Belaj et al., 2012; Domínguez-García et al., 2012; Trujillo et al., 2014). In general, there were more clusters containing similar associations between individuals than different ones. Indeed, in this context, HRM could be considered a very sensitive approach.

HRM analyses has been previously used in other species (Simko, 2016). Yet, to the best of our knowledge, this is the first report of such genotyping technology with olive tree. Besides, a large WOGBC number of samples from many different countries were analyzed. Therefore, such approach is proposed as a new tool to evaluate genetic diversity and structure in olive tree. It only requires qRT-PCR and *in silico* analyses, reducing time and costs in experimental workflows. Furthermore, this approach can be applied to any collection of samples with known sequences, as it is the case for WOGBC (Ganopoulos et al., 2011).

3.6. Conclusions

Nowadays, molecular markers are broadly used to identify, classify, further analyze, manage, and protect genetic diversity. Notwithstanding, their developments are challenging, as they require previous genetic-background knowledge of studied species. Additionally, experimental work, cost and time may be very high (Ovesná et al., 2014). Fortunately, unlike other classical molecular-marker approaches, HRM analyses allow the detection of small genomic variations (such as SNP or indels), exhibiting a high-resolution power, as described by its own name. Overall, it has been shown that HRM analyses are an effective approach to evaluate genetic diversity of the WOGBC. To the best of our knowledge, this is the first report of HRM to genotype olive tree, including WOGBC accessions from different countries. Such new tool generated relevant information for further studies, including the same or other molecular markers. Such an approach can be used in the future to identify traits of agronomical interest.

3.7. References

ALBA, R., Z. FEI, P. PAYTON, Y. LIU, S.L. MOORE, P. DEBBIE, J. COHN, M. D'ASCENZO, J. S. GORDON, J. K. C. ROSE, G. MARTIN, S. D. TANKSLEY, M. BOUZAYEN, M. M. JAHN, and J. GIOVANNONI. 2004. ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. *The Plant Journal: For Cell and Molecular Biology* 39: 697–714.

ANGIOLILLO, A., M. MENCUCCINI, and L. BALDONI. 1999. Olive genetic diversity assessed using amplified fragment length polymorphisms. *Theoretical and Applied Genetics* 98: 411–421.

BALDONI, L., and A. BELAJ. 2009. Olive. *In* J. Vollmann, and I. Rajcan [eds.], Oil crops, Handbook of Plant Breeding, 397–421. Springer New York.

BARTOLINI, G., R. PETRUCCELLI, and F. AND A.O. OF THE U. NATIONS. 2002. Classification, origin, diffusion and history of the Olive. *Food and Agriculture Organization* (FAO).

BEAUCHAMP, G.K., R.S.J. KEAST, D. MOREL, J. LIN, J. PIKA, Q. HAN, C.-H. LEE, A. B. SMITH, and P. A. S. BRESLIN. 2005. Phytochemistry: Ibuprofen-like activity in extra-virgin olive oil. *Nature* 437: 45–46.

BELAJ, A., M. DEL C. DOMINGUEZ-GARCÍA, S.G. ATIENZA, N.M. URDÍROZ, R.D. LA ROSA, Z. SATOVIC, A. MARTÍN, A. KILIAN, I. TRUJILLO, V. VALPUESTA, and C. DEL RÍO. 2012. Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genetics & Genomes* 8: 365–378.

BESNARD, G., P. BARADAT, and A. BERVILLÉ. 2001. Genetic relationships in the olive (*Olea europaea* L.) reflect multilocal selection of cultivars. *Theoretical and Applied Genetics* 102: 251–258.

BRACCI, T., M. BUSCONI, C. FOGHER, and L. SEBASTIANI. 2011. Molecular studies in olive (*Olea europaea* L.): overview on DNA markers applications and recent advances in genome analysis. *Plant Cell Reports* 30: 449–462.

BUSCONI, M., L. SEBASTIANI, and C. FOGHER. 2006. Development of SCAR markers for germplasm characterisation in olive tree (*Olea europea* L.). *Molecular Breeding* 17: 59–68.

BUSHAKRA, J.M., M.J. STEPHENS, A.N. ATMADJAJA, K.S. LEWERS, V.V. SYMONDS, J.A. UDALL, D. CHAGNÉ, E. J. BUCKS. and E. GARDINER. 2012. Construction of black (*Rubus occidentalis*) and red (*R. idaeus*) raspberry linkage maps and their comparison to the genomes of strawberry, apple, and peach. *Theoretical and Applied Genetics* 125: 311–327.

CABALLERO J.M., C. DEL RÍO, C. NAVARRO, M.D. GARCIA-FERNANDEZ, J. MORALES, M. HERMOSO, L.A. DEL OLMO, F. LOPEZ, F. CERA, G. RUIZ. 2005. Ensayos comparativos en Andalucía. *In* Rallo L, Barranco D, Caballero J, Martín A, Del Río C, Tous J, Trujillo I [eds]. Variedades de olivo en España, vol 2, MAPA. Ediciones Mundi- Prensa and COI, Sevilla, pp 383–394

CARUSO, M., G. DISTEFANO, D. PIETRO PAOLO, S. LA MALFA, G. RUSSO, A. GENTILE, and G.R. RECUPERO. 2014. High resolution melting analysis for early identification of citrus hybrids: A reliable tool to overcome the limitations of morphological markers and assist rootstock breeding. *Scientia Horticulturae* 180: 199–206.

CRONQUIST, A. 1981. An integrated system of classification of flowering plants. Columbia University Press.

CRUZ, F., I. JULCA, J. GÓMEZ-GARRIDO, D. LOSKA, M. MARCET-HOUBEN, E. CANO, B. GALÁN, L. FRIAS, P. RIBECA, S. DERDAK, M. GUT, M. SÁNCHEZ-FERNÁNDEZ, J. L. GARCÍA, I. G. GUT, P. VARGAS, T. S. ALIOTO and T. GABALDÓN. 2016. Genome sequence of the olive tree, *Olea europaea*. *GigaScience* 5: 29.

DIGHE, A.S., K. JANGID, J.M. GONZALEZ, V.J. PIDIYAR, M.S. PATOLE, D.R. RANADE, and Y.S. SHOUCHE. 2004. Comparison of 16S rRNA gene sequences of genus Methanobrevibacter. *BMC Microbiology* 4: 20.

DISTEFANO, G., S. LA MALFA, S. CURRÒ, G. LAS CASAS, A. WÜNSCH, and A. GENTILE. 2015. HRM analysis of chloroplast and mitochondrial DNA revealed additional genetic variability in Prunus. *Scientia Horticulturae* 197: 124–129.

DISTEFANO, G., S.L. MALFA, A. GENTILE, and S. B. WU. 2013. EST-SNP genotyping of citrus species using high-resolution melting curve analysis. *Tree Genetics & Genomes* 9: 1271–1281.

DOMÍNGUEZ-GARCÍA, M.C., A. BELAJ, R. DE LA ROSA, Z. SATOVIC, K. HELLER-USZYNSKA, A. KILIAN, A. MARTÍN, and S.G. ATIENZA. 2012. Development of DArT markers in olive (*Olea europaea* L.) and usefulness in variability studies and genome mapping. *Scientia Horticulturae* 136: 50–60.

DORADO, G., S. GÁLVEZ, H. BUDAK, T. UNVER, and P. HERNÁNDEZ. 2015. Nucleic-acid sequencing. *In* Caplan M [ed]. Reference Module in Biomedical Sciences. Biochemistry, Cell Biology and Molecular Biology. Elsevier Amsterdam.

FABBRI, A., M. LAMBARDI, and Y. OZDEN-TOKATLI. 2009. Olive breeding. *In* S. M. Jain, and P. M. Priyadarshan [eds.], Breeding plantation tree crops: Tropical species, 423–465. Springer New York.

FELSENSTEIN. 1989. PHYLIP - phylogeny Inference Package. *Cladistics* 1989, 5: 164-166

GANOPOULOS, I., A. ARGIRIOU, and A. TSAFTARIS. 2011. Microsatellite high resolution melting (SSR-HRM) analysis for authenticity testing of protected designation of origin (PDO) sweet cherry products. *Food Control* 22: 532–541.

GOMES, S., P. MARTINS-LOPES, J. LIMA-BRITO, J. MEIRINHOS, J. LOPES, A. MARTINS, and H. GUEDES-PINTO. 2008. Evidence for clonal variation in "Verdeal-Transmontana" olive using RAPD, ISSR and SSR markers. *The Journal of Horticultural Science and Biotechnology* 83: 395–400.

GREEN, P.S. 2002. A Revision of *Olea* L. (*Oleaceae*). *Kew Bulletin* 57: 91–140.

HERNANDEZ, P., R. DE LA ROSA, L. RALLO, A. MARTIN, and G. DORADO. 2001. First evidence of a retrotransposon-like element in olive (*Olea europaea*): implications

in plant variety identification by SCAR-marker development. *Theoretical and Applied Genetics* 102: 1082–1087.

KUMAR, S., G. STECHER, and K. TAMURA. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33: 1870–1874.

LAS CASAS, G., F. SCOLLO, G. DISTEFANO, A. CONTINELLA, A. GENTILE, and S. LA MALFA. 2014. Molecular characterization of olive (*Olea europaea* L.) Sicilian cultivars using SSR markers. *Biochemical Systematics and Ecology* 57: 15–19.

LIU, K., and S.V. MUSE. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics (Oxford, England)* 21: 2128–2129.

MACKAY, J.F., C.D. WRIGHT, and R.G. BONFIGLIOLI. 2008. A new approach to varietal identification in plants by microsatellite high resolution melting analysis: application to the verification of grapevine and olive cultivars. *Plant Methods* 4: 8.

MARIOTTI, R., N.G.M. CULTRERA, S. MOUSAVI, F. BAGLIVO, M. ROSSI, E. ALBERTINI, F. ALAGNA, F. CARBONE, G. PERROTTA, and L. BALDONI. 2016. Development, evaluation, and validation of new EST-SSR markers in olive (*Olea europaea* L.). *Tree Genetics & Genomes* 12: 120.

MINISTERIO DE AGRICULTURA Y PESCA, ALIMENTACIÓN Y MEDIO AMBIENTE. 2015. Encuesta sobre Superficies y Rendimientos Cultivos (ESYRCE). Available at: http://www.mapama.gob.es/es/estadistica/temas/estadisticas-agrarias/agricultura/esyrce/# [Accessed February 11, 2017].

MOHAMMADI, S., and B. PRASANNA. 2003. Analysis of Genetic Diversity in Crop Plants—Salient Statistical Tools and Considerations. *Crop Science Society of America* 43: 1235–1248.

MONTEMURRO, C., M.M. MIAZZI, A. PASQUALONE, V. FANELLI, W. SABETTA, and V. DI RIENZO. 2015. Traceability of PDO olive oil “Terra di Bari” using high resolution melting. *Journal of Chemistry* 2015: e496986.

MULEO, R., M.C. COLAO, D. MIANO, M. CIRILLI, M.C. INTRIERI, L. BALDONI, and E. RUGINI. 2009. Mutation scanning and genotyping by high-resolution DNA melting analysis in olive germplasm. *Genome* 52: 252–260.

MUÑOZ-MÉRIDA, A., J.J. GONZÁLEZ-PLAZA, A. CAÑADA, A.M. BLANCO, M. DEL C. GARCÍA-LÓPEZ, J.M. RODRÍGUEZ, L. PEDROLA, M. D. SICARDO, M. L. HERNÁNDEZ, R. DE LA ROSA, A. BELAJ, M. GIL-BORJA, F. LUQUE, J. M. MARTÍNEZ-RIVAS, D. G. PISANO, O. TRELLES, V. VALPUESTA, and C. R. BEUZÓN. 2013. *De novo* assembly and functional annotation of the Olive (*Olea europaea*) transcriptome. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 20: 93–108.

MURRAY, M.G., and W.F. THOMPSON. 1980. Rapid isolation of high molecular-weight plant DNA. *Nucleic Acids Research* 8: 4321–4325.

NEI, M., F. TAJIMA, and Y. TATENO. 1983. Accuracy of estimated phylogenetic trees from molecular-data .2. Gene-frequency data. *Journal of Molecular Evolution* 19: 153–170.

NYBOM, H., and I. BARTISH. 2000. Effects of life history traits and sampling strategies on genetic diversity estimates obtained with RAPD markers in plants. *Perspectives in Plant Ecology, Evolution and Systematics* 3: 93–114.

OVESNÁ, J., L. LEIŠOVÁ-SVOBODOVÁ, and L. KUČERA. 2014. Microsatellite analysis indicates the specific genetic basis of Czech bolting garlic. *Czech Journal of Genetics and Plant Breeding* 50: 226–234.

PASQUALONE, A., V. DI RIENZO, W. SABETTA, V. FANELLI, C. SUMMO, V.M. PARADISO, C. MONTEMURRO, and F. CAPONIO. 2016. Chemical and molecular characterization of crude oil obtained by olive-pomace recentrifugation. *Journal of Chemistry* 2016: e4347207.

PASQUALONE, A., V.D. RIENZO, M.M. MIAZZI, V. FANELLI, F. CAPONIO, and C. MONTEMURRO. 2015. High resolution melting analysis of DNA microsatellites in olive pastes and virgin olive oils obtained by talc addition. *European Journal of Lipid Science and Technology* 117: 2044–2048.

PÉREZ-JIMÉNEZ, F., J. RUANO, P. PEREZ-MARTINEZ, F. LOPEZ-SEGURA, and J. LOPEZ-MIRANDA. 2007. The influence of olive oil on human health: not a question of fat alone. *Molecular Nutrition & Food Research* 51: 1199–1208.

RODRIGUES, F., F. B. PIMENTEL, and MBPP. OLIVEIRA. 2015. Olive by-products: Challenge application in cosmetic industry. *Industrial Crops and Products* 70: 116–124.

SERVILI, M., B. SORDINI, S. ESPOSTO, S. URBANI, G. VENEZIANI, I. DI MAIO, R. SELVAGGINI, and A. TATICCHI. 2013. Biological Activities of Phenolic Compounds of Extra Virgin Olive Oil. *Antioxidants* 3: 1–23.

SIMKO, I. 2016. High-Resolution DNA Melting Analysis in Plant Research. *Trends in Plant Science* 21: 528–537.

SIMKO, I., I. EUJAYL, and T.J.L. VAN HINTUM. 2012. Empirical evaluation of DArT, SNP, and SSR marker-systems for genotyping, clustering, and assigning sugar beet hybrid varieties into populations. *Plant Science* 184: 54–62.

TRUJILLO, I., M.A. OJEDA, N.M. URDIROZ, D. POTTER, D. BARRANCO, L. RALLO, and C.M. DIEZ. 2014. Identification of the Worldwide Olive Germplasm Bank of Córdoba (Spain) using SSR and morphological markers. *Tree Genetics & Genomes* 10: 141–155.

VIETINA, M., C. AGRIMONTI, and N. MARMIROLI. 2013. Detection of plant oil DNA using high resolution melting (HRM) post PCR analysis: a tool for disclosure of olive oil adulteration. *Food Chemistry* 141: 3820–3826.

XANTHOPOULOU, A., I. GANOPOULOS, G. KOUBOURIS, A. TSAFTARIS, C. SERGENDANI, A. KALIVAS, and P. MADESIS. 2014. Microsatellite high-resolution melting (SSR-HRM) analysis for genotyping and molecular characterization of an *Olea europaea* germplasm collection. *Plant Genetic Resources* 12: 273–277.

ZHAO, W., J. CHUNG, G. LEE, K. MA, H. KIM, K. KIM, I. CHUNG, J. K. LEE, N. S. KIM, S. M. KIM, and Y. J. PARK. 2010. Molecular genetic diversity and population

structure of a selected core set in garlic and its relatives using novel SSR markers. *Plant Breeding* 130: 46–54.

3.8. Supplementary Material

Supplementary Material 3.1. Country distribution of olive-tree varieties analyzed.

Numbering	Cultivation area	Reference	Numbering	Cultivation area	Reference
1	Algeria	ARG01	43	Spain	SPA16
2	Chile	CHI01	44	Spain	SPA17
3	Croatia	CRO01	45	Spain	SPA18
4	Croatia	CRO02	46	Spain	SPA19
5	Croatia	CRO03	47	Spain	SPA20
6	Croatia	CRO04	48	Spain	SPA21
7	Egypt	EGY01	49	Spain	SPA22
8	France	FRA01	50	Spain	SPA23
9	France	FRA02	51	Spain	SPA24
10	France	FRA03	52	Spain	SPA25
11	France	FRA04	53	Spain	SPA26
12	Greece	GRE01	54	Spain	SPA27
13	Greece	GRE02	55	Spain	SPA28
14	Israel	ISR01	56	Spain	SPA29
15	Israel	ISR02	57	Spain	SPA30
16	Italy	ITA01	58	Spain	SPA31
17	Italy	ITA02	59	Spain	SPA32
18	Italy	ITA03	60	Spain	SPA33
19	Italy	ITA04	61	Spain	SPA34
20	Italy	ITA05	62	Spain	SPA35
21	Italy	ITA06	63	Spain	SPA36
22	Lebanon	LEB01	64	Spain	SPA37
23	Morocco	MOR01	65	Spain	SPA38
24	Morocco	MOR02	66	Spain	SPA39
25	Morocco	MOR03	67	Spain	SPA40
26	Portugal	POR01	68	Spain	SPA41
27	Portugal	POR02	69	Spain	SPA42
28	Spain	SPA01	70	Syria	SYR01
29	Spain	SPA02	71	Syria	SYR02
30	Spain	SPA03	72	Syria	SYR03
31	Spain	SPA04	73	Syria	SYR04
32	Spain	SPA05	74	Tunisia	TUN01
33	Spain	SPA06	75	Tunisia	TUN02
34	Spain	SPA07	76	Tunisia	TUN03
35	Spain	SPA08	77	Tunisia	TUN04
36	Spain	SPA09	78	Turkey	TUR01
37	Spain	SPA10	79	Turkey	TUR02
38	Spain	SPA11	80	Turkey	TUR03
39	Spain	SPA12	81	Turkey	TUR04
40	Spain	SPA13	82	Turkey	TUR05
41	Spain	SPA14	83	Turkey	TUR06
42	Spain	SPA15			

GENERAL CONCLUSIONS /

CONCLUSIONES GENERALES

1. In general, DArTseq data for garlic is consistent with prior passport data.

2. The number of accessions of the analyzed garlic germplasm-bank has been significantly reduced, identifying redundant ones, generating a core collection, and, therefore, reducing space and maintenance costs.

3. Some DArTseq reads are associated to Gene Ontology processes.

4. DArTseq technology is a cost-effective method to perform high-throughput genetic diversity analyses in the absence of reference genome, even with species like garlic, with huge and expected complex genome.

5. DArTseq genotyping results should be useful for garlic-bank curators to identify, manage and protect biodiversity, as well as for plant breeders, to improve garlic varieties.

6. To the best of our knowledge, this is the first high-throughput genotyping-by-sequencing in garlic by DArTseq technology.

7. High-Resolution Melting analysis provides a “closed-tube” technique suitable for olive-tree genotyping.

1. En general, los datos de DArTseq de ajo son consistentes con la información previa de los datos de pasaporte.

2. El número de entradas analizadas del banco de germoplasma de ajo ha sido reducido significativamente, identificado entradas redundantes, generando una colección nuclear y, por tanto, este análisis puede utilizarse para reducir los costes de espacio y mantenimiento del banco.

3. Una fracción de las lecturas de DArTseq generadas pueden ser asociadas a procesos de Ontología Génica.

4. La tecnología DArTseq es un método rentable para realizar análisis de diversidad genética de alto rendimiento en ausencia de un genoma de referencia, incluso con especies como el ajo, con un genoma esperado complejo y enorme.

5. El genotipado mediante DArTseq es una herramienta útil para los conservadores de bancos para identificar, gestionar y proteger la biodiversidad, además de para los mejoradores de plantas, para mejorar variedades de ajo.

6. Hasta donde sabemos, este es el primer genotipado por secuenciación de alto rendimiento en ajo usando la tecnología DArTseq.

7. El análisis HRM ("High-Resolution Melting") aporta una técnica de "tubo cerrado" adecuada para el genotipado del olivo.