

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Mining SNPs and linkage analysis in *Cynara cardunculus*

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/155007> since 2016-07-14T13:55:38Z

Publisher:

Springer Netherlands

Published version:

DOI:10.1007/978-94-007-7572-5_22

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)



UNIVERSITÀ DEGLI STUDI DI TORINO

This is an author version of the contribution published on the book:

*Genomics of Plant Genetic Resources, Volume 1: managing, sequencing and mining
genetic resources - DOI 10.1007/978-94-007-7572-5_22*

The definitive version is available at:

<http://www.springer.com/it/book/9789400775718>

1 Mining SNPs and linkage analysis in *Cynara cardunculus*

2 Sergio Lanteri, Alberto Acquadro, Davide Scaglione, Ezio Portis

3 University of Torino, DIVAPRA, Via Leonardo da Vinci 44 – 10095 - Grugliasco (TO) – Italy; sergio.lanteri@unito.it

5 Abstract

6 *Cynara cardunculus* L., a member of the *Asteraceae* family, is a diploid ($2n=34$) outcrossing
7 perennial species native to the Mediterranean basin. It includes globe artichoke (var. *scolymus* L.),
8 which today is grown as vegetable all over the world, cultivated cardoon (var. *altilis* DC), locally
9 grown in Southern European countries, and their progenitor wild cardoon (var. *sylvestris* Lam). The
10 species is also a valuable source of pharmaceutical compounds, and is exploitable for the
11 production of lignocellulosic biomass as well as oil from seed, the latter being suitable for both
12 edible and bio-fuel end-uses.

13 By crossing a non spiny globe artichoke genotype (female parent) with selected genotypes of the
14 tree botanical taxa, we generated three F1 segregating progenies from which genetic maps, based
15 on the two-way pseudo test cross strategy, have been developed. From the globe artichoke and
16 cultivated cardoon genetic maps a reference SSR-based consensus map was constructed, which
17 consists of 227 loci (217 SSRs and ten SNPs) assembled into 17 major linkage groups. To further
18 saturate the *C. cardunculus* maps we recently applied NGS (next generation sequencing)
19 technologies for mining a wide set of SNPs (single nucleotide polymorphism). Based on Illumina
20 sequencing of gDNA RAD (restriction associated DNA) tags of three mapping parents (e.g. non
21 spiny globe artichoke, cultivated and wild cardoon), we generated ~19.000 genomic contigs (mean
22 312 bp) and ~17.000 SNPs (density 1/139 bp). Side by side, the transcriptome of the same mapping
23 parents was sequenced by using a 454 platform, and raw data *de novo* assembled and annotated to
24 generate the first reference transcriptome of the species (38,726 unigenes, 32.7 Mbp).

25 The 454 reads, together with Illumina paired ends (PEs) from further eight *C. cardunculus*
26 genotypes were aligned on the reference contig set, and ~195.000 SNPs were called (density
27 1/169bp in coding regions). The two workflows led to produce a massive set of SNPs in *C.*
28 *cardunculus*, and made possible create an extensive gene catalogue as a valuable resource for
29 upcoming genomic and genetic studies.

30
31
32
33
34

35 **1.1 The *Cynara cardunculus* complex**

36 *Cynara cardunculus* L. is native to the Mediterranean Basin and includes three botanical taxa: the
37 globe artichoke (var. *scolymus*), the cultivated cardoon (var. *altilis*) and the wild cardoon [var.
38 *sylvestris* (Lamk) Fiori]. The three forms are fully cross-compatible with one another, and form
39 fertile hybrids (Basnizki and Zohary 1994). Reproductive barriers separate the *C. cardunculus*
40 complex from the other *Cynara* species (Rottenberg et al. 1996). The crosses between *C.*
41 *cardunculus* and the wild species *C. syriaca*, *C. algarbiensis*, *C. baetica* and *C. humilis* do all
42 produce few seeds, although the hybrids are generally sterile; the wild species are therefore
43 regarded as members of the secondary wild gene pool of *C. cardunculus* (Rottenberg and Zohary
44 2005). On both morphological (Wiklund 1992) and cytogenetic (Rottenberg et al. 1996) grounds,
45 the closest of the wild species to the cultivated complex is *C. syriaca*. The monophyly and
46 evolution of the *Cynara* spp. have been investigated by sequence comparisons between various ITS
47 (internal transcribed spacer) and ETS (external transcribed spacer) regions (Robba et al. 2005;
48 Sonnante et al. 2007) leading to the suggestion that the *cardunculus* complex is more differentiated
49 and evolved than the other wild species.

50 Molecular (Lanteri et al. 2004; Acquadro et al. 2005), cytogenetic and isozyme (Rottenberg
51 et al. 1996) studies have confirmed that wild cardoon is the ancestor of both the domesticate globe
52 artichoke and cultivated cardoon, which evolved independently under the influence of distinct
53 anthropogenic selection criteria. The earliest report of the presence of *C. cardunculus* in Sicily and
54 Greece dates back to Theophrastus (371–287 BCE), while in 77 CE, the Roman naturalist Pliny the
55 Elder mentioned its use for medicinal purposes; however, little is known either of the process of
56 domestication or the subsequent diversification of the two taxa. It has been assumed that before
57 globe artichoke was selected, cardoon was cultivated for its fleshy stems and roots, which were
58 considered a delicacy by the ancient Greeks and Romans (Portis et al. 2005a; Portis et al. 2005b).
59 On the other hand, the best guess is that the globe artichoke was domesticated and transformed into
60 the plant that we know today, most probably between 800 and 1500 CE in family or monastery
61 gardens. Recently, by assessing the AFLP pattern of genetic diversity of a collection of Sicilian
62 globe artichoke landraces, which have been cultivated for a number of centuries by local farmers,
63 one landrace was identified which appears to represent an early stage of the domestication process,
64 suggesting Sicily as one of the possible centre of globe artichoke domestication (Mauro et al. 2008).

65 Globe artichoke contributes significantly to the Mediterranean agricultural economy, with an
66 annual production of about 750 metric tons (MT) from over 80,000 ha of cultivated land and with
67 an annual turnover exceeding US \$ 500 million. Italy is the leading world producer (480 MT/year,
68 FAOSTAT 2010), followed by Egypt and Spain. Globe artichoke cultivation is increasing in South

69 America and the United States, and more recently also in China. The prime globe artichoke product
70 consists of the immature inflorescence (heads of capitula), which can be consumed in fresh, canned
71 or frozen form. Each plant produces a number of capitula, the largest of which (the main capitulum)
72 merges from the apex of the central stem, while the smaller ones are produced on lateral branches.

73 Italy has the richest globe artichoke primary cultivated “gene pool” and harbours many
74 distinct clonal varietal groups, best adapted to local environments. On the basis of harvest time,
75 varietal types can be classified as early, producing heads from autumn to spring, and late, producing
76 heads from early to late spring. On the basis of capitulum characters, cultivated germplasm has been
77 classified into four main groups: (1) the Spinosi group, containing types with long sharp spines on
78 bracts and leaves; (2) the Violetti group, with medium-sized, violet-coloured and less spiny heads;
79 (3) the Romaneschi group, with spherical or subspherical non-spiny heads; (4) the Catanesi group,
80 with relatively small, elongated and non-spiny heads. The classification based on head is in
81 consistent agreement with the one obtained by assessing the AFLP genetic variation in a wide
82 collection of 84 varietal types grown worldwide, indicating that the cultivated morphotypes play an
83 important role in determining variation within the cultivated globe artichoke germplasm (Lanteri et
84 al. 2004). Although in recent years some seed (achenes)-propagated varieties have been introduced,
85 but vegetative propagation, by means of basal and lateral offshoots (either semi-dormant or actively
86 growing), or stump pieces, has been adopted for centuries, and it is still largely prevalent in most of
87 the varietal types and local landraces. Due to the limited selection adopted by farmers on the mother
88 plants used for vegetative propagation, as well as mutations occurred over time, the populations at
89 present in cultivation are multiclonal and characterized by a wide range of within population genetic
90 variation (Lanteri et al. 2001; Portis et al. 2005c).

91 The cultivated cardoon (*C. cardunculus* var. *altilis* DC) is usually raised from seed and
92 handled as annual crop; its cultivation is much less widespread than that of the globe artichoke and
93 the crop remains of regional importance in Spain, Italy and the south of France, where it is used in
94 traditional dishes. The edible parts of the plant are the fleshy stems which are typically collected in
95 late autumn-early winter and often, before collection are tied together, wrapped in straw, and/or
96 buried for about three weeks in order to accentuate the flavour. A study based on SSR and AFLP
97 profiling of the most widely grown Italian and Spanish local varieties showed that they form two
98 separate gene-pools and that a considerable level of within variety variation is present (Portis et al.
99 2005b).

100 The wild cardoon is a robust thistle distributed over the west and central part of the
101 Mediterranean basin (Portugal to west Turkey) as well as Canary Islands; in post Columbian time it
102 colonized some part of the New World and has spread as a weed in Argentina and California

103 (Marushia and Holt 2006). Its flowers have been used for centuries in the Iberian Peninsula for
104 manufacturing of ovine and caprine milk cheese (Sousa and Malcata 1996; Barbagallo et al. 2007)
105 and its small and thorny capitula are sometimes sold in local markets in Sicily (Ierna and
106 Mauromicale 2010).

107 108 **1.2 Uses of globe artichoke and cardoon other than for human food**

109 *C. cardunculus* has long been known to represent a valuable source of biopharmaceutical
110 compounds (Slanina et al. 1993; Wagenbreth 1996; Sevcikova et al. 2002; Wang et al. 2003). Roots
111 and rhizomes, used also for brew or infusion, provide a source of inulin, a demonstrated enhancer of
112 the human intestinal flora, while leaves and heads represent one of the richest natural source of
113 compounds originating from the metabolism of phenylpropanoids, with caffeoylquinic acids and
114 flavonoids as major components. *C. cardunculus* extracts influence glucose and lipid metabolism
115 (Blumenthal et al. 2000) and were reported to be effective in increasing the feeling of satiety in
116 overweight subjects (Rondanelli et al. 2011); in various pharmacological test systems it has been
117 demonstrated that they (i) protect proteins lipids and DNA from oxidative damage from free
118 radicals (Gebhardt 1997; Brown and Rice-Evans 1998; Perez-Garcia et al. 2000), (ii) inhibit
119 cholesterol biosynthesis and contribute to the prevention of atherosclerosis and other vascular
120 disorders (Kraft 1997; Brown and Rice-Evans 1998; Gebhardt 1998; Pittlern and Ernst 1998;
121 Matsui et al. 2006; Bundy et al. 2008). Furthermore, it has been demonstrated that *C. cardunculus*
122 extracts inhibit HIV integrase, a key player in HIV replication and its insertion into host DNA
123 (McDougall et al. 1998; Slanina et al. 2001), possess apoptotic properties (Miccadei et al. 2008) and
124 exert antibacterial activity (Martino et al. 1999).

125 The composition of the globe artichoke phenolic fraction includes four mono-caffeoylquinic
126 isomers, six dicaffeoylquinic isomers, six flavonoid glycosides, and at least seven anthocyanins
127 (Lattanzio et al. 2009). The genes involved in the biosynthesis of the mono-caffeoylquinic acid
128 (chlorogenic acid) have been identified as well as their regulation under UV-C stress (Comino et al.
129 2007, 2009; De Paolis et al. 2008; Moglia et al. 2009; Menin et al. 2010; Sonnante et al. 2010).
130 Conversely, the biosynthetic pathway leading to di-caffeoylquinic acids is a matter of debate
131 (Villegas and Kojima 1986; Hoffmann et al. 2003; Niggeweg et al. 2004).

132 The characteristic bitterness of both globe artichoke and cultivated cardoon is mainly due to
133 the presence of sesquiterpene lactones (STLs), of which the two major representatives are
134 cynaropicrin and, at lower concentration, grosheimin and its derivatives (Schneider and Thiele
135 1974; Cravotto et al. 2005). Cynaropicrin, like many sesquiterpenes lactones, has various medicinal
136 activities (Shimoda et al. 2003; Cho et al. 2004; Schinor et al. 2004; Emendorfer et al. 2005; Ishida

137 et al. 2010) among which cytotoxicity against several types of cancer cells (Yasukawa et al. 2010).
138 In globe artichoke a germacrene A synthase, involved in the first step of STLs biosynthesis, has
139 been recently isolated, functionally characterized and mapped (Menin et al. 2012).

140
141 *C. cardunculus* has great potential as a source of renewable energy, thanks to its
142 productivity of lignocellulosic biomass. The caloric value of the three *C. cardunculus* taxa is
143 analogous, however cultivated cardoon has the highest biomass yield, which can reach up to ~
144 19t/ha dry matter with an energy value ~ 17 MJ/kg (Angelini et al. 2009; Ierna and Mauromicale
145 2010; Portis et al. 2010; Ierna et al. 2012). The species has been also identified as a candidate for
146 the production of seed oil which is suitable for both comestible and bio-fuel end-uses. Seed yield in
147 cardoon is about 2 t/ha and up to 0.8 t/ha in globe artichoke (at 5% w/v moisture), from about 25 to
148 30% of which is oil of good alimentary quality (Foti et al. 1999) due to its high and well balanced
149 content of oleic and linoleic acids, its low content of free acids, peroxides, saturated and linoleic
150 acids and a favourable α -tocopherol content (Maccarone et al. 1999), while the seed material left
151 after oil extraction can be used as a component of animal feed.

151

152 **1.3 Linkage analyses: state of the art**

153 The genome organization of *C. cardunculus* ($2n=2x=34$; haploid genome size ~1.08 Gbp),
154 unlike other species belonging to the Asteraceae family (e.g. sunflower, lettuce and chicory),
155 remains largely unexplored. The species is an out-breeder, and is characteristically highly
156 heterozygous. Its marked level of inbreeding depression inhibits the use of backcross, F_2 or
157 recombinant inbred populations for mapping purposes. As haploid induction - via either
158 androgenesis or gynogenesis - has not yet been achieved (Motzo and Deidda 1993; Chatelet et al.
159 2005; Stamigna et al. 2005), no possibility is presently available to generate doubled haploid
160 populations. Thus, genetic mapping in the species has relied on a double pseudo-testcross approach
161 (Grattapaglia and Sederoff 1994), in which segregating F_1 progeny are derived from a cross
162 between two heterozygous individuals.

163 The first genetic maps of *C. cardunculus* were provided by Lanteri et al. (2006), and based
164 on a cross between two genotypes of globe artichoke, namely the varietal types ‘Romanesco C3’ (a
165 late-maturing non-spiny type used as female) and ‘Spinoso di Palermo’ (an early-maturing spiny
166 type used as male). This population was genotyped using a number of PCR-based marker platforms,
167 resulting in a ~1300 cM female map consisting of 204 loci, divided into 18 linkage groups (LGs),
168 and a ~1200 cM male map comprising 180 loci and 17 LGs. The two maps shared 78 loci, which
169 allowed for the alignment of 16 of the LGs. The maps have since been extended by the inclusion of
170 three genes involved in the synthesis of caffeoylquinic acid (Comino et al. 2009; Moglia et al. 2009)

171 and a number of microsatellite loci, of which 19 were represented in both maps (Acquadro et al.
172 2009).

173 New maps have lately been generated from a set of F₁ progeny involving the cross between
174 the same female parent as previously ('Romanesco C3') and the cultivated cardoon genotype
175 'Altilis 41' (Portis et al. 2009a). The cultivated cardoon map comprised 177 loci, subdivided into 17
176 LGs and spanning just over 1000 cM, while the globe artichoke one featured 326 loci arranged into
177 20 LGs, spanning ~1500 cM with a mean inter-marker distance of ~ 4.5 cM. A set of 84 loci shared
178 between this 'Romanesco C3' map and the previously developed one (Lanteri et al. 2006) allowed
179 for map alignment and the definition of 17 homologous LGs, corresponding to the haploid
180 chromosome number of the species. Later on, the maps have been integrated with the inclusion of
181 all the genes involved in the synthesis of caffeoylquinic acids known in the species (Menin et al.
182 2010).

183 Since more markers were needed to saturate the maps, a further wide set of SSR markers
184 was developed from ESTs (expressed sequence tags) of globe artichoke, made available by the
185 Composite Genome Project (CGP; <http://compgenomics.ucdavis.edu/>). Using a custom
186 bioinformatic pipeline, 36,321 ESTs were assembled into 19,055 unigenes (6,621 contigs and
187 12,434 singletons), annotated, and mined for perfect SSRs. Over 4,000 potential EST-SSR loci,
188 lying within some 3,300 genes (one SSR per 3.6 kbp) have been identified, and PCR primers for the
189 amplification of more than 2,000 of these have been designed. In a test of a sample of 300 of these
190 assays, over half proved to be informative between the parents of the available mapping populations
191 (Scaglione et al. 2009). As a result, a large number of these EST-SSR loci have been integrated into
192 the globe artichoke and cultivated cardoon maps (Portis et al. 2012) and, more recently, exploited
193 for genetic mapping in a population obtained by crossing globe artichoke with wild cardoon
194 (Sonnante et al. 2011). The integration of 139 EST-SSR loci has significantly improved the
195 resolution and accuracy of the 'Romanesco C3' and 'Altilis 41' maps. The female map was built
196 with 473 loci spanning 1.544 cM with a mean inter-marker distance of 3.4 cM, corresponding to a
197 3.8% increase in length over the earlier map, but in a ~ 28% decrease in the mean inter-marker
198 distance. The male map consisted of 273 loci spanning 1486 cM, with a mean inter-marker distance
199 of 5.4 cM, representing a marked increase in both length (+42%) and number of loci (+50%),
200 together with a minor decrease in the mean inter-marker distance (-5%). The two maps shared 66
201 codominant loci (64 SSRs and two SNPs), which allowed for the alignment of all the 'Romanesco
202 C3' with the 'Altilis 41' LGs. Following alignment a consensus linkage map based exclusively on
203 microsatellite and SNP markers (as depicted in Figure 1) was constructed (Portis et al. 2009b). The
204 consensus map is shown in Figure 2; it comprised 227 loci (217 SSRs and ten SNPs targeting genes

205 involved in the synthesis of caffeoylquinic acids) arranged into 20 LGs (LOD threshold > 6.0). The
206 consensus map length was 1068.0 cM, with a mean inter-marker spacing of 5.2 cM . The length of
207 LGs varied from 4.0 to 113.7 cM (mean 62.8 cM), with the largest LG containing 36 loci. Lowering
208 the LOD threshold to 5.0 resulted in the merging of three pairs of LG, thereby reducing the overall
209 number to 17, corresponding to the haploid chromosome complement of the species. The majority
210 of the LGs contained a mixture of ‘Romanesco C3’, ‘Altilis 41’ and shared co-dominant markers,
211 with only four (LG_9, 13, 14 and 7) carrying shared loci and markers only present in the
212 ‘Romanesco C3’ map.

213 Putative functions have been deduced for SSR markers derived from ESTs using homology
214 searches with public protein databases. Annotation of mapped loci was performed via BlastX search
215 as well as InterPro scan and GO categorisation made it possible to tag some biological functions. A
216 set of 17 EST-SSR markers were annotated with GO terms involved in the ‘response to stimulus’
217 (Table 1), five and eight of which were derived from transcripts related to response to cold and salt
218 stress, respectively. As an example, the marker CyEM-42, developed from the contig
219 CL4773Contig1 (Scaglione et al. 2009) and mapped on LG_12 of the SSR-based consensus map,
220 showed high aminoacidic similarity (81%) with the protein kinase PBS1 of *Arabidopsis*. To
221 consider reliable orthology, a reciprocal tblastx analysis against the whole EST collection currently
222 available for *C. cardunculus* was performed, and no better alignment than that of contig CL4773
223 was detected. PBS1 was found to work as R gene against the bacterial pathogen *Pseudomonas*
224 *syringae*, where its cleavage, operated by the pathogens’ effector AvrPphB, triggers the signalling
225 cascade, generating the host response (Shao et al. 2002). *Pseudomonas* spp. together with other
226 endophytic bacteria may affect globe artichoke plants both in field and during micropropagation
227 (Penalver et al. 1994) and the CyEM-42 may be likely considered a reliable marker for tagging a
228 bacterial resistance trait in the species. On the whole, these EST-SSR markers may be defined as
229 functional markers with the potential to target polymorphisms in gene responsible for traits of
230 interest; they can be also particularly useful for constructing comparative framework maps with
231 other Asteraceae, giving the possibility to amplify ortholog genes and provide anchor loci.

232 This SSR-based consensus map of *C. cardunculus* is based on a robust marker platform of
233 SSRs and a few gene-based SNP loci. It is expected that the further positioning of markers within
234 target regions will provide key tools for marker-assisted breeding programs as well as the necessary
235 framework to exploit mapping data obtained from diverse populations. At present, ~ 200 of the loci
236 on the consensus map (about 88%) are sited within genic sequence, presenting some opportunity to
237 identify candidate genes for particular traits within the species.

238

239 **1.4 SNP mining**

240 The first set of SNP (single nucleotide polymorphism) markers available for the species has been
241 developed on genes involved in the synthesis of caffeoylquinic acid, as above reported. The allelic
242 forms of globe artichoke acyltransferases HCT, HQT (Comino et al. 2007; Comino et al. 2009) and
243 the hydroxylase C3'H (Moglia et al. 2009), were analysed in the two globe artichoke parental
244 genotypes ('Romanesco C3' and 'Spinoso di Palermo') of the first genetic maps (Lanteri et al.
245 2006) and SNPs were identified. SNP genotyping of the F₁ progeny was carried out with the tetra-
246 primers ARMS-PCR method (Ye et al. 2001; Chiapparino et al. 2004). A further SNP set has been
247 later on developed by Menin et al. (2010) on three acyltransferases and on the *C4H*, *4CL* and
248 *MYB12* genes, identified by an *in silico* scan of the globe artichoke unigene set assembled by
249 Scaglione et al. (2009). Gene homologues were re-sequenced in the parental genotypes (globe
250 artichoke 'Romanesco C3' and cultivated cardoon 'Atilis 41') of the genetic maps developed by
251 Portis et al. (2009a) and genes successfully mapped.

252 Recent advances in next-generation DNA sequencing technologies have made possible the
253 development of high-throughput SNP genotyping platforms, that allow for the simultaneous
254 interrogation of thousands of SNPs. Such resources have the potential to facilitate the rapid
255 development of high-density genetic maps, and to enable genome-wide association studies as well
256 as molecular breeding approaches in a variety of taxa (Bachlava et al. 2012). Thousands of SNPs
257 have been recently developed in *C. cardunculus* by Next-Generation Sequencing (NGS) technology
258 using two complementary approaches (Figure 3):

259 1) genomic RAD (Restriction-site Associated DNA) tag sequencing (Miller et al. 2007) in
260 combination with the Illumina Genome Analyzer sequencing device (Baird et al. 2008) of three
261 genotypes (globe artichoke, cultivated cardoon and wild cardoon) that were crossed for developing
262 F₁ mapping populations (Scaglione et al. 2012a);

263 2) transcriptome sequencing, via 454 and Illumina technologies, of the same three genotypes
264 plus eight, five of which were globe artichoke, two cultivated and one wild cardoon (Scaglione et
265 al. 2012b). Alongside, a functional characterisation and annotation of the obtained sequence set was
266 performed. These SNPs represent a one-stop resource to produce a dense *C. cardunculus* genetic
267 map via high-throughput genotyping technologies.

268

269 **1.4.1 Genomic SNP mining**

270 The recently developed restriction-site associated DNA (RAD) approach (Box 1) has been
271 combined with the Illumina DNA sequencing platform to enable the rapid and mass discovery of
272 SNP markers. Three genomic RAD libraries were obtained from the three *C. cardunculus* genotypes

273 belonging to the three taxa of the species and parents of two mapping populations. The first
274 mapping population is an F₁ progeny involving the cross between globe artichoke ('Romanesco
275 C3', female parent) and cultivated cardoon (genotype 'Atilis 41') (Portis et al. 2009a). The second
276 one is an F₁ progeny involving the cross between the same female parent as previously and the wild
277 cardoon (genotype 'Creta4') (Lanteri et al. 2011).

278

279 **1.4.1.1 RAD tag sequencing and *de novo* assembly**

280 The RAD-seq exercise produced 9.7 million reads (19.4 million Pair End - PE), equivalent to ~ 1
281 Gbp of sequence. The distribution of reads was uneven across the three DNA samples, with 1.2
282 million reads achieved for globe artichoke, 2.6 million for cultivated cardoon and 5.9 million for
283 wild cardoon; the latter, being the largest set, was chosen as the basis for *de novo* contigs assembly.
284 The assembly procedure created 19,061 reference genomic contigs, spanning 6.11 Mbp (with N50 =
285 321 bp and a mean a contig length of 312 bp). The GC content was ~ 37.4% which is similar to that
286 of many dicots species (Jaillon et al. 2007) and represents the first survey on the base composition
287 of the *C. cardunculus* genome.

288

289 **1.4.1.2 RAD tag annotation**

290 The contig sequences characterisation was conducted using the BlastX algorithm and it resulted in
291 the annotation of 5,335 contigs (28.0%). Regardless of the genome-wide RAD sampling, a
292 noteworthy part of the annotated sequences might be represented by coding regions, since a
293 methylation-sensitive enzyme (*Pst*I) was used to produce the RAD-tag libraries (Palmer et al.
294 2003), although the rather short length of the RAD contigs made difficult to distinguish between
295 putative genes and pseudogenes. Enzyme codes were retrieved for 1,327 contigs, defining a unique
296 set of 313 putative enzymatic activities, which were mapped onto KEGG reference pathways
297 (<http://www.genome.jp/kegg/>). The remaining portion of the contig set (72%) was not attributed to
298 any known sequence, likely due to the RAD contigs shortness.

299 The transposable DNA element footprints detected, using RepeatMasker software (v3.2.9;
300 <http://www.repeatmasker.org>) implemented with the RMBlast algorithm, and adopting the
301 *Viridiplantae* repeats as reference, accounted for a 0.2% of the sequence, while 1.2% of the
302 sequences derived from LTR retroelements, including Ty/Copia-like (0.8%) and Gypsy-like (0.2%).
303 This quantification of transposable element abundance could have been underestimated, but these
304 data represents a useful snapshot of relative abundance of each different mobile element class in *C.*
305 *cardunculus*.

306

307 **1.4.1.3 SNP calling**

308 The PE sequences generated for each mapping parent were aligned using the reference contig set as
309 a scaffold. In total, ~ 33,000 sequence variants were detected, including 1,520 short indels,
310 distributed over 12,068 contigs. The overall SNP frequency was estimated to be 5.6 per 1,000
311 nucleotides, a level which is almost equal to that found in the non-coding regions of the *V. vinifera*
312 genome (5.5 per 1,000 nucleotides; Velasco et al. 2007) and very similar to that observed in *Citrus*
313 spp. ESTs (6.1 per 1,000 nucleotides; Jiang et al. 2010). A subset of ~ 17,400 SNPs was obtained
314 considering allelic variant which were informative for both mapping populations (16,727 SNPs, and
315 723 1-2nt indels) distributed over 7,478 contigs.

316 Since *C. cardunculus* is highly heterozygous, SNPs were categorized as intra- or inter-
317 varietal, where the former also represents the heterozygous state of the analysed genotype. The two
318 types were not exclusive, therefore heterozygous SNPs present in one sample could be found in
319 both heterozygous or homozygous states in other genotypes. The number of heterozygous SNP loci
320 was 1,235 in the globe artichoke, 2,868 in the cultivated cardoon and 5,069 in the wild cardoon
321 mapping parents (Figure 4). Heterozygous SNPs are of key importance for mapping studies since
322 for the linkage analysis a two-way pseudo-testcross approach, based on a segregant F₁ progeny, was
323 adopted. In this sense, a key parameter for the successful isolation of such useful SNP markers was
324 the sequencing coverage.

325

326 **1.4.2 Transcriptomic SNP mining**

327 A total of eleven *C. cardunculus* EST libraries were produced and after the normalisation
328 procedures, they were separately sequenced. Three libraries, deriving from the three mapping
329 parents (Table 2), were sequenced with the 454 Titanium (Roche) to produce a reference
330 transcriptome. Eight libraries, set up from five globe artichoke accessions, two cultivated cardoon
331 and one wild cardoon genotypes (Table 3), were sequenced using the Illumina GAIIx platform, in
332 order to highly increase the total SNP calling amount.

333

334 **1.4.2.1 EST sequencing and *de novo* assembly**

335 The outcome of 454-based cDNA sequencing of the three mapping parents generated some 1.7 M
336 reads of overall length 695 Mb, which were reduced to 692 Mb after a post-sequencing filtering.
337 The mean read length was equal to 392 bp (Table 2). cDNA libraries of other eight genotypes
338 (Table 3) were sequenced using a GAIIx platform (Illumina) producing 6.9 Gbp of raw data (46.4
339 M paired-end reads) with a mean of 5.8 M reads per accession. The data set was reduced to 6.7 Gbp
340 following the removal of adaptor sequences and other contaminants, and it was further reduced to

341 6.2 Gbp after quality trimming. For the *de novo* assembly process only the 454 reads were
342 considered, while the Illumina data were simply adopted to increase the efficiency of the SNP
343 mining process.

344 The assembly of 454 reads was achieved by a two-tier approach using the MIRA assembler
345 ver.3.2.0 (Chevreux et al. 2004). In a first step, each individual sample was assembled
346 independently. The process generated 37,622 contigs for ‘Romanesco C3’, 40,130 contigs for
347 ‘Altilis 41’, and 42,837 contigs for ‘Creta 4’ with N50 contig lengths of 834 bp, 761 bp and 772 bp,
348 and mean coverage levels of 7.31, 8.45 and 9.17X, respectively. For the ‘Romanesco C3’
349 assembly, a subset of 11,276 contigs resulted from the incorporation of a prior set of 28,641 Sanger
350 ESTs (www.ncbi.nlm.nih.gov/dbEST). Then, after contaminant removal by BLASTX analysis, the
351 three datasets were merged into a set of 38,726 contigs. This “reference” assembly spanned 32.7
352 Mbp and had a GC content of 42.1%. The mean contig length was 844.3 bp (N50: 951 bp).

353 A second assembly phase was carried out by merging at least two *taxon*-derived contigs
354 from the first phase, and 20,469 contigs were generated. They consisted of a subset with a mean
355 length of 1054 bp, while 5,375, 6,669 and 6,213 remained as single *taxon*-derived contigs of var.
356 *scolymus*, var. *altilis* and var. *sylvestris*, respectively.

357

358 1.4.2.3 Sequence analysis and functional annotation

359 The sequence reads were assembled into 38,726 reference transcripts, which were successfully
360 annotated, using the Blast2GO pipeline, by gene ontology terms via Blast and InterPro analyses.
361 Enzymes were tagged on KEGG's reference pathways (www.genome.jp/kegg/), including primary
362 and secondary metabolisms. On the whole, 16,419 enzyme codes were retrieved (12,449 transcripts)
363 and subsequently mapped onto KEGG's pathways. The sample of *C. cardunculus* enzymes
364 consisted of 1,133 unique enzyme codes distributed across 147 pathways. To provide an example,
365 by analyzing the whole transcriptome complement, a subset of 71 enzymatic activities involved in
366 phenylpropanoid synthesis were identified; 21 of these were annotated at varying levels of
367 redundancy in the core phenylpropanoid pathway (KEGG's map: 00940), in which the synthesis of
368 caffeoylquinic and di-caffeoylquinic acids (CQAs and dCQAs) takes place (Figure 5).

369 Transcriptional factor function was assigned to 1,398 transcripts, scattered across 67
370 families, while 316 sequences were tagged as candidate Resistance Gene Analogs (RGAs). Each
371 sequence was scanned for the presence of recognition sites for known plant miRNAs. In total, target
372 annealing sites for 302 miRNAs were located in 1,043 transcripts, which mainly belong to the

373 categories: “defense response” and “programmed cell death/apoptosis”, “reproduction”,
374 “development of anatomical structure”, “photosynthesis”, “transmembrane receptor activity” and
375 “transcription factor activity”. The 454-based assembly included non-nuclear transcripts. The *C.*
376 *cardunculus* chloroplast genes identification was based on similarity to those of lettuce and
377 sunflower (Timme et al. 2007) leading to the categorization of 137 contigs, of which 80 were
378 putatively assigned the chloroplast genome. Similarly, the grapevine (*Vitis vinifera*) mitochondrial
379 genes (Goremykin et al. 2009) aided in the identification of 52 *C. cardunculus* contigs, which were
380 putatively attributed to the mitochondrial genome.

381 To estimate the transcriptome representation and its gene-level redundancy (e.g. splicing
382 variants), two different approaches were adopted. Using the *A. thaliana* gene content, the 454
383 sequencing output was predicted to be assembled in a total of 29.3 Mbp, distributed in 24,064
384 unigenes (average length of 1,216 bp) which covered 96% of the transcriptome. Alternatively, the
385 final contig set (38,000) was clustered by collapsing gene variants (e.g. alternative splicing), which
386 generated a set of 29,830 unigenes that represents a *bona fide* estimation of the gene content of *C.*
387 *cardunculus*. Data suggest that 23% of splicing variants could be present in the transcriptome
388 assembly.

389

390 **1.4.2.4 Read mapping and SNP calling**

391 About 1.5 M of the 454-derived reads were aligned to the reference contig set (38,726 contigs).
392 This number was reduced to ~ 1.0 M by removing those that showed more than one unique
393 alignment, thereby lowering the risk of false SNP calls due to misalignment of paralog-derived
394 reads or to redundancy resulting from splicing variants. The same procedure was repeated for the
395 Illumina-derived reads, producing an alignment of ~ 60 M paired ends. Resolving paired ends
396 reduced this to a set of ~ 21 M reads.

397

398

399 An assembly based on about 35 M sequences was generated by merging the 454 and Illumina
400 sequence datasets, resulting in a median reference transcriptome coverage of 96X with 26,990
401 reference contigs containing at least 20 mapped reads. Reliable SNPs (Bayesian probability >95%)
402 were detected at 195,400 sites across the set of 11 accessions. The average SNP frequency was
403 calculated at one per 167 bp, with a mean of five per contig. Each SNP site was interrogated by
404 scoring for the presence of at least one accession-specific sequence. Sequence information was
405 available from an average of nine accessions per SNP site, and a core subset of 57,125 SNPs
406 showed coverage from all the samples. The merging of the Illumina-derived reads (eight

407 accessions) with 454-generated reads substantially increased the number of parent-specific SNPs
408 that were identified (Figure 6).

409 SNP frequency in the *C. cardunculus* transcriptome appears to be comparable to that found
410 in the heterozygous grapevine whole genome sequence (Velasco et al. 2007) and among *Citrus* ssp.
411 ESTs (Jiang et al. 2010). Overall, SNPs were most frequent in 3'-UTR (one per 126 bp), followed
412 by the CDS (one per 169 bp), and the 5'-UTR (one per 265 bp). Within the UTRs, the frequency
413 also matched that obtained in tomato expressed sequence (Jimenez-Gomez and Maloof 2009), while
414 it was markedly different to that present in the coding region (~ 2 per kb). This discrepancy may
415 reflect either the greater tolerance by the heterozygous state of non-synonymous substitutions, or
416 merely is an ascertainment bias due to the analysis of a larger germplasm panel which also included
417 accessions of a wild relative.

418 In *C. cardunculus*, as previously pointed out, the presence of intra-accession allelic variation
419 is of particular interest. As expected by their shallower coverage, the 454-derived sequences
420 produced a somewhat lower frequency of SNPs with successful heterozygous SNP calling (Figure
421 7). 'Altilis 41' was relatively the least heterozygous of the accessions (17,570 loci), as has been
422 observed previously (Portis et al. 2005b; Portis et al. 2009a), while 'Romanesco Zorzi' was the
423 most heterozygous (43,387 loci), followed by 'Violetto di Chioggia' (41,824 loci). 'Imperial Star'
424 had the lowest ratio of heterozygous variants among globe artichoke genotypes (13.5%), which
425 likely reflects its development from crosses among less genetically differentiated genotypes.

426

427 **1.4.3 Conclusions**

428 The second generation technologies provide high sequencing throughputs at significantly reduced
429 costs if compared to Sanger. These platforms are currently employed for large-scale SNP discovery
430 projects and, for medium-scale projects, they have been frequently applied in combination with
431 reduced-complexity libraries, targeting genomic subsets.

432 One such method, aimed at decreasing the sample complexity, is to build up a genomic library, with
433 a reduced locus representation including only a subset of sequences generated by restriction
434 enzymes, which cut at frequent intervals throughout the genome. The generation of an SNP set can
435 be achieved through the deep-sequencing of such libraries and the comparison between allelic
436 variants can identify thousands of SNPs.

437 The recent RAD (Baird et al. 2008) approach is focused on the targeting of a discrete
438 number of genomic regions adjacent to specific restriction sites, and it can effectively reduce the
439 number of the fragments to be sequenced in a given complex genome. This strategy (see Box 1)
440 represents a promising experimental scheme in term of costs and technical simplicity and, so far,

441 has been successfully adopted for SNP discovery in many plant and animal species (Davey et al.
442 2011).

443 An alternative approach is to focus onto the transcriptome deep-sequencing, which reduces
444 the representation of low information-content repetitive sequences in species possessing a large
445 genome and/or without a finished genome sequencing project. An EST library can lead to identify a
446 large number of genetic loci and targeting SNPs in coding sequences. This kind of library represents
447 a one-stop resource useful for many downstream applications and to address many biological
448 questions in plant science. It can aid the identification of genes underlying phenotypes of interest
449 through the development of expression arrays or provide thousands of loci as a source of potential
450 markers for QTL mapping applications and population genomic studies.

451 The two experimental workflows led to produce a massive set of SNPs in *C. cardunculus*, and
452 made possible create an extensive gene catalogue, as a valuable resource for upcoming genomic and
453 genetic studies. Both approaches have proven to be efficient for SNP mining, although
454 characterized by peculiarities and limitations which deserve to be considered in view of specific
455 research targets. In *C. cardunculus* the EST sequencing approach generated a set of reference
456 coding sequences spanning 32.7 Mbps, establishing a ‘general gene catalog’ of 38,726 as *bona fide*
457 representation of the transcriptome. In contrast the RAD-tag sequencing approach permitted to
458 sequence 6.0 Mbps separated in lesser and shorter number of contigs (~ 19,000; 28% of which were
459 annotated as CDS-like). The number of SNPs was higher for EST than for RAD-tag approach
460 (195,000 vs. 34,000); nevertheless, the SNP frequency observed in the two pipelines were
461 somewhat comparable (5.9 vs. 5.6 per 1,000 nt). The RAD-tags data revealed to be extremely
462 informative to preliminary survey the repetitive DNA component of the *C. cardunculus* genome,
463 and allowed us to make some inferences regarding the contribution of DNA methylation in
464 inhibiting its expansion (Scaglione et al. 2012a).

465 From the standpoint of costs, RAD technology was attempted with a great technical
466 simplicity and a low cost/time expense. The cDNA library setting up was indeed more complex for
467 both the need of standardization/normalisation procedures and some extra enzymatic steps required,
468 however, side by side, its sequencing output provided a better picture of the globe artichoke coding
469 genome. Bearing in mind a future in which the globe artichoke genome will be completely
470 sequenced and publicly available, the genomic RAD approach may represent one of the most
471 feasible and cheap strategy for accomplishing affordable targeted re-sequencing projects. It is also
472 likely that the increasingly lowering of sequencing costs will make the scientific community to
473 converge towards new approaches of ‘genotyping-by-sequencing’. This scheme proceeds to explore

474 all the nucleotidic positions of a genome in a single experiment, and will permit an integration of
475 mapping and sequencing steps, likely bypassing many costly physical mapping procedures.

476 The combination of two NGS platforms (454 FLX Titanium - Roche and GAIIx - Illumina)
477 for the extensive characterization of the genome and transcriptome of *C. cardunculus*, has proven to
478 be a highly reliable tools for SNP discovery. Overall, the availability of such a large number of
479 sequence-based markers, in a format allowing for high-throughput genotyping, offers opportunities
480 to developed a high-density genetic map and association mapping studies aimed at correlating
481 molecular polymorphisms with variation in phenotypic traits, as well as for molecular breeding
482 approaches in a species which has multiple end-uses such as food, nutraceuticals and bioenergy.
483 The high number of mined SNPs represents also an excellent resource for evolutionary genetic
484 studies in cultivated forms and their wild relative as well as for comparative genetic mapping
485 studies aimed at understanding patterns of genome rearrangement between *C. cardunculus* and
486 related species.

487

488 **1.5 Acknowledgements**

489 We wish to thank:

- 490 - Loren H. Rieseberg, Steven J. Knapp and Zhao Lai (Compositae Genome Project) for founding
491 the RAD tag and transcriptome sequencing within the U.S. National Science Foundation grants
492 “Comparative genomics of phenotypic variation in the Compositae (DBI-0820451)”
493 - Giovanni Mauromicale and Rosario P. Mauro (Dipartimento di Scienze delle Produzioni Agrarie e
494 Alimentari (DISPA) - Sez. Scienze Agronomiche, University of Catania) for the development and
495 maintenance in field of the mapping progenies.

496

497 **1.6 References**

- 498 Acquadro A, Portis E, Lee D et al. (2005) Development and characterization of microsatellite
499 markers in *Cynara cardunculus* L. *Genome* 48:217-225
- 500 Acquadro A, Lanteri S, Scaglione D et al. (2009) Genetic mapping and annotation of genomic
501 microsatellites isolated from globe artichoke. *Theor Appl Genet* 118:1573-1587
- 502 Angelini LG, Ceccarini L, Di Nasso NNO, Bonari E (2009) Long-term evaluation of biomass
503 production and quality of two cardoon (*Cynara cardunculus* L.) cultivars for energy use.
504 *Biomass Bioenerg* 33:810-816
- 505 Bachlava E, Taylor CA, Tang S et al. (2012) SNP discovery and development of a high-density
506 genotyping array for sunflower. *PLoS one* 7:e29814
- 507 Baird N, Etter P, Atwood T et al. (2008) Rapid SNP discovery and genetic mapping using
508 sequenced RAD markers. *PLoS one* 3:e3376
- 509 Barbagallo RN, Chisari M, Spagna G et al. (2007) Caseinolytic activity expression in flowers of
510 *Cynara cardunculus* L., *Acta Hort* 730:195-199
- 511 Blumenthal M, Goldberg A, Brinckmann J (2000) Artichoke leaf. eds. *Herbal Medicine: Expanded*
512 *Commission E Monographs, Integrative Medicine Communications* Boston, MA, 10-210-12
- 513 Brown J, Rice-Evans C (1998) Luteolin-rich artichoke extract protects low density lipoprotein from
514 oxidation in vitro. *Free Rad Res* 29:247-255
- 515 Bundy R, Walker A, Middleton R et al. (2008) Artichoke leaf extract (*Cynara scolymus*) reduces
516 plasma cholesterol in otherwise healthy hypercholesterolemic adults: A randomized, double
517 blind placebo controlled trial. *Phytomedicine* 15:668-675
- 518 Chatelet P, Stamigna C, Thomas G (2005) Early development from isolated microspores of *Cynara*
519 *cardunculus* var. *scolymus* (L.) Fiori. *Acta Hort* 681:375-380
- 520 Chevreux B, Pfisterer T, Drescher B et al. (2004) Using the miraEST assembler for reliable and
521 automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res*
522 14:1147-1159
- 523 Chiapparino E, Lee D, Donini P (2004) Genotyping single nucleotide polymorphisms in barley by
524 tetra-primer ARMS-PCR. *Genome* 47:414-420
- 525 Cho J, Kim A, Jung J et al. (2004) Cytotoxic and pro-apoptotic activities of cynaropicrin, a
526 sesquiterpene lactone, on the viability of leukocyte cancer cell lines. *Eur J Pharmacol* 492:85-
527 94
- 528 Comino C, Lanteri S, Portis E et al. (2007) Isolation and functional characterization of a cDNA
529 coding a hydroxycinnamoyltransferase involved in phenylpropanoid biosynthesis in *Cynara*
530 *cardunculus* L. *BMC Plant Biol* 7:14
- 531 Comino C, Hehn A, Moglia A et al. (2009) The isolation and mapping of a novel
532 hydroxycinnamoyltransferase in the globe artichoke chlorogenic acid pathway. *BMC Plant Biol*
533 9:30
- 534 Cravotto G, Nano G, Binello A et al. (2005) Chemical and biological modification of cynaropicrin
535 and grosheimin: a structure-bitterness relationship study. *J Sc Food Agric* 85:1757-1764
- 536 Davey JW, Hohenlohe PA, Etter PD et al. (2011) Genome-wide genetic marker discovery and
537 genotyping using next-generation sequencing. *Nat Rev Genet* 12: 499-510
- 538 De Paolis A, Pignone D, Morgese A et al. (2008) Characterization and differential expression
539 analysis of artichoke phenylalanine ammonia-lyase-coding sequences. *Phys Plant* 132:33-43
- 540 Emendorfer F, Emendorfer F, Bellato F et al. (2005) Antispasmodic activity of fractions and
541 cynaropocrin from *Cynara scolymus* on guinea-pig ileum. *Biol Pharm Bull* 28:902-904
- 542 Foti S, Mauromicale G, Raccuia S et al. (1999) Possible alternative utilization of *Cynara* spp. I.
543 Biomass, grain yield and chemical composition of grain. *Ind Crop Prods* 10:219-228

- 544 Gebhardt R (1997) Antioxidative and protective properties of extracts from leaves of the artichoke
545 (*Cynara scolymus* L) against hydroperoxide-induced oxidative stress in cultured rat
546 hepatocytes. *Tox Appl Pharm* 144:279-286
- 547 Gebhardt R (1998) Inhibition of cholesterol biosynthesis in primary cultured rat hepatocytes by
548 artichoke (*Cynara scolymus* L.) extracts. *J Pharm Exp Therapy* 286:1122-1128
- 549 Goremykin V, Salamini F, Velasco R, Viola R (2009) Mitochondrial DNA of *Vitis vinifera* and the
550 issue of rampant horizontal gene transfer. *Mol Biol Evol* 26:99-110
- 551 Grattapaglia D, Sederoff R. (1994) Genetic-linkage maps of *Eucalyptus grandis* and *Eucalyptus*
552 *urophylla* using a pseudo-testcross - mapping strategy and RAPD markers. *Genetics* 137:1121-
553 1137
- 554 Hoffmann L, Maury S, Martz F et al. (2003) Purification, cloning, and properties of an
555 acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid
556 metabolism. *J Biol Chem* 278:95-103
- 557 Ierna A, Mauromicale G. (2010) *Cynara cardunculus* L. genotypes as a crop for energy purposes in
558 a Mediterranean environment. *Biomass Bioenerg* 34:754-760
- 559 Ierna A, Mauro RP, Mauromicale G (2012) Biomass, grain and energy yield in *Cynara cardunculus*
560 L. as affected by fertilization, genotype and harvest time. *Biomass Bioenerg* 36:404-410
- 561 Ishida K, Kojima R, Tsuboi M et al. (2010) Effects of artichoke leaf extract on acute gastric
562 mucosal injury in rats. *Biol Pharm Bull* 33:223-229
- 563 Jaillon O, Aury J, Noel B et al. (2007) The grapevine genome sequence suggests ancestral
564 hexaploidization in major angiosperm phyla. *Nature* 449:463-467
- 565 Jiang D, Ye Q, Wang F, Cao L (2010) The mining of *Citrus* EST-SNP and its application in cultivar
566 discrimination. *Agricultural Sciences in China* 9:79-190
- 567 Jimenez-Gomez J, Maloof J (2009) Sequence diversity in three tomato species: SNPs, markers, and
568 molecular evolution. *BMC Plant Biol* 9:85
- 569 Kraft K (1997) Artichoke leaf extract - Recent findings reflecting effects on lipid metabolism, liver
570 and gastrointestinal tracts. *Phytomed* 4:369-378
- 571 Lanteri S, Di Leo I, Ledda L et al. (2001) RAPD variation within and among populations of globe
572 artichoke cultivar 'Spinoso sardo'. *Plant Breed* 120:243-246
- 573 Lanteri S, Saba E, Cadinu M et al. (2004) Amplified fragment length polymorphism for genetic
574 diversity assessment in globe artichoke. *Theor Appl Genet* 108:1534-1544
- 575 Lanteri S, Acquadro A, Comino C et al. (2006) A first linkage map of globe artichoke (*Cynara*
576 *cardunculus* var. *scolymus* L.) based on AFLP, S-SAP, M-AFLP and microsatellite markers.
577 *Theor Appl Genet* 112:1532-1542
- 578 Lanteri S, Portis E, Acquadro A et al. (2011) Morphology and SSR fingerprinting of newly
579 developed *Cynara cardunculus* genotypes exploitable as ornamentals. *Euphytica* 184:311-321
- 580 Lattanzio V, Kroon PA, Linsalata V, Cardinali A (2009) Globe artichoke: A functional food and
581 source of nutraceutical ingredients. *J Func Foods* 1:131-144
- 582 Maccarone E, Fallico B, Fanella F et al. (1999) Possible alternative utilization of *Cynara* spp. II.
583 Chemical characterization of their grain oil. *Ind Crops Prod* 1:229-237
- 584 Martino V, Caffini N, Phillipson J et al. (1999) Identification and characterization of antimicrobial
585 components in leaf extracts of globe artichoke (*Cynara scolymus* L.). *Acta Hort* 501:111-114
- 586 Marushia RG, Holt JS (2006) The effects of habitat on dispersal patterns of an invasive thistle,
587 *Cynara cardunculus*. *Biol Invas* 8:577-593
- 588 Matsui T, Ogunwande IA et al. (2006) Anti-hyperglycemic potential of natural products. *Mini Rev*
589 *Med Chem* 6:349-356

- 590 Mauro R, Portis E, Acquadro A et al. (2008) Genetic diversity of globe artichoke landraces from
591 Sicilian small-holdings: implications for evolution and domestication of the species. *Cons*
592 *Genet* 10:431-440
- 593 McDougall B, King P, Wu B et al. (1998) Dicafeoylquinic and dicafeoyltartaric acids are selective
594 inhibitors of human immunodeficiency virus type 1 integrase. *Antimicrob Agents Chemother*
595 42:140-146
- 596 Menin B, Comino C, Moglia A et al. (2010) Identification and mapping of genes related to
597 caffeoylquinic acid synthesis in *Cynara cardunculus* L. *Plant Sc* 179:338-347
- 598 Menin B, Comino C, Portis E et al. (2012) Genetic mapping and characterization of the globe
599 artichoke (+)-germacrene A synthase gene, encoding the first dedicated enzyme for
600 biosynthesis of the bitter sesquiterpene lactone cynaropicrin. *Plant Sc* 190:1-8
- 601 Miccadei S, Di Venere D, Cardinali A et al. (2008) Antioxidative and apoptotic properties of
602 polyphenolic extracts from edible part of artichoke (*Cynara scolymus* L.) on cultured rat
603 hepatocytes and on human hepatoma cells. *Nutr Cancer* 60:276-283
- 604 Miller M, Dunham J, Amores A et al. (2007) Rapid and cost-effective polymorphism identification
605 and genotyping using restriction site associated DNA (RAD) markers. *Genome Res* 17:240-248
- 606 Moglia A, Comino C, Portis E et al. (2009) Isolation and mapping of a C3'H gene (CYP98A49)
607 from globe artichoke, and its expression upon UV-C stress. *Plant Cell Rep* 28:963-974
- 608 Motzo R, Deidda M (1993) Anther and ovule culture in globe artichoke. *J. Genet. Breed.*, 47:263-
609 266
- 610 Niggeweg R, Michael A, Martin C (2004) Engineering plants with increased levels of the
611 antioxidant chlorogenic acid. *Nat Biotech* 22:746-754
- 612 Palmer LE, Rabinowicz PD, O'Shaughnessy AL et al. (2003) Maize genome sequencing by
613 methylation filtrations. *Science* 302:2115-2117
- 614 Penalver R, Duranvila N, Lopez MM (1994) Characterization and pathogenicity of bacteria from
615 shoot tips of the globe artichoke (*Cynara scolymus*). *Ann Appl Biol* 125:501-513
- 616 Perez-Garcia F, Adzet T, Canigual S (2000) Activity of artichoke leaf extract on reactive oxygen
617 species in human leukocytes. *Free Radic Res* 33:661-665
- 618 Pittlern M, Ernst E. (1998) Artichoke leaf extract for serum cholesterol reduction. *Perfusion* 11:338-
619 340
- 620 Portis E, Acquadro A, Comino C et al. (2005a) Genetic structure of island populations of wild
621 cardoon [*Cynara cardunculus* L. var. *sylvestris* (Lamk) Fiori] detected by AFLPs and SSRs.
622 *Plant Sc* 169:199-210
- 623 Portis E, Barchi L, Acquadro A et al. (2005b). Genetic diversity assessment in cultivated cardoon
624 by AFLP (amplified fragment length polymorphism) and microsatellite markers. *Plant Breed*
625 124:299-304
- 626 Portis E, Mauromicale G, Barchi L et al. (2005c). Population structure and genetic variation in
627 autochthonous globe artichoke germplasm from Sicily Island. *Plant Sc* 168:1591-1598
- 628 Portis E, Mauromicale G, Mauro R et al. (2009a) Construction of a reference molecular linkage
629 map of globe artichoke (*Cynara cardunculus* var. *scolymus*). *Theor Appl Genet* 120:59-70
- 630 Portis E, Acquadro A, Scaglione D et al. (2009b). Construction of an SSR-based linkage map for
631 *Cynara cardunculus*. *Proceeding of the 8th Plant Genomics European Meeting*, p 142
- 632 Portis E, Acquadro A, Longo A, Mauro R, Mauromicale G, Lanteri S. (2010) Potentiality of *Cynara*
633 *cardunculus* L. as energy crop. *J Biotech* 150:S165-S166
- 634 Portis E, Scaglione D, Acquadro A et al. (2012) Genetic mapping and identification of QTL for
635 earliness in the globe artichoke/cultivated cardoon complex. *BMC Res Notes* 5:252

- 636 Robba L, Carine M, Russell S, Raimondo F (2005) The monophyly and evolution of *Cynara* L.
637 (Asteraceae) *sensu lato*: evidence from the Internal Transcribed Spacer region of nrDNA. *Plant*
638 *Syst Evol* 253:53-64
- 639 Rondanelli M, Giacosa A, Orsini F et al. (2011) Appetite Control and Glycaemia Reduction in
640 Overweight Subjects treated with a Combination of Two Highly Standardized Extracts from
641 *Phaseolus vulgaris* and *Cynara scolymus*. *Phytother Res* 25:1275-1282
- 642 Rottenberg A, Zohary D, Nevo E (1996) Isozyme relationships between cultivated artichoke and the
643 wild relatives. *Genetic Resources and Crop Evolution* 43:59-62
- 644 Rottenberg A, Zohary D (2005) Wild genetic resources of cultivated artichoke. *Acta Horti* 681:307-
645 311
- 646 Scaglione D, Acquadro A, Portis E et al. (2009) Ontology and diversity of transcript-associated
647 microsatellites mined from a globe artichoke EST database. *BMC Genomics* 10:454
- 648 Scaglione D, Acquadro A, Portis E et al. (2012a). RAD tag sequencing as a source of SNP markers
649 in *Cynara cardunculus* L. *BMC Genomics* 13:3
- 650 Scaglione D, Lanteri S, Acquadro A et al. (2012b). Large-scale transcriptome characterization and
651 mass discovery of SNPs in globe artichoke and its related taxa. *Plant Biotech J* 10:956-969.
- 652 Schinor E, Salvador M, Ito I et al. (2004) Trypanocidal and antimicrobial activities of *Moquinia*
653 *kingii*. *Phytomedicine* 11:224-229
- 654 Schneider G, Thiele K 1974. Die Verteilung des Bitter-stoffes Cynaropicrin in der Artischocke.
655 *Planta Medica* 26:174-183
- 656 Sevcikova P, Glatz Z, Slanina J (2002) Analysis of artichoke (*Cynara cardunculus* L.) extract by
657 means of micellar electrokinetic capillary chromatography. *Electrophoresis* 23:249-252
- 658 Shao F, Merritt P, Bao Z et al. (2002) A *Yersinia* effector and a *Pseudomonas* avirulence protein
659 define a family of cysteine proteases functioning in bacterial pathogenesis. *Cell* 109:575-588
- 660 Shimoda H, Ninomiya K, Nishida N et al. (2003) Anti-hyperlipidemic Sesquiterpenes and new
661 sesquiterpene glycosides from the leaves of artichoke (*Cynara scolymus* L.): Structure
662 requirement and mode of action. *Bioorg Med Chem Lett* 3:223-228.
- 663 Slanina J, Taborska E, Bochorakova H et al. (2001) New and facile method of preparation of the
664 anti-HIV-1 agent, 1,3-dicaffeoylquinic acid. *Tetrahedron Letters* 42:3383-3385
- 665 Slanina J, Taborska E, Musil P (1993) Determination of cynarine in the decoctions of the artichoke
666 (*Cynara cardunculus* L.) by the HPLC method. *Cesko-SloV Farm*, 42:265-268
- 667 Sonnante G, D'Amore R, Blanco E et al. (2010) Novel hydroxycinnamoyl-coenzyme a quinate
668 transferase genes from artichoke are involved in the synthesis of chlorogenic acid. *Plant*
669 *Physiol* 153:1224-1238
- 670 Sonnante G, Pignone D, Hammer K (2007) The domestication of artichoke and cardoon: From
671 roman times to the genomic age. *Ann Bot* 100:1095-1100
- 672 Sonnante G, Gatto A, Morgese A et al. (2011) Genetic map of artichoke × wild cardoon: toward a
673 consensus map for *Cynara cardunculus*. *Theor Appl Genet* 123:1215-1229
- 674 Sousa MJ, Malcata FX (1996) Influence of pasteurization of milk and addition of starter cultures on
675 protein breakdown in ovine cheeses manufactured with extracts from flowers of *Cynara*
676 *cardunculus*. *Food Chem* 57:549-556
- 677 Stamigna C, Saccardo F, Pandozy G et al. (2005) In vitro mutagenesis of globe artichoke (cv.
678 Romanesco). *Acta Hort* 681:403-410
- 679 Timme R, Kuehl J, Boore J, Jansen R (2007) A comparative analysis of the *Lactuca* and *Helianthus*
680 (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared
681 repeats. *Am J Bot* 94:302-312
- 682 Velasco R, Zharkikh A, Troglio M et al. (2007) A High Quality Draft Consensus Sequence of the
683 Genome of a Heterozygous Grapevine Variety. *PLoS one* 2:e1326

- 684 Villegas R, Kojima M (1986) Purification and characterization of Hydroxycinnamoyl D-Glucose:
685 quinate hydroxycinnamoyl transferase in the root of sweet potato, *Ipomoea batatas* Lam. J Biol
686 Chem 261:8729-8733
- 687 Wagenbreth D (1996) Evaluation of artichoke cultivars for growing and pharmaceutical use. Beitr
688 Zuchtungsforsch 2:400-403
- 689 Wang M, Simon J, Aviles I et al. (2003) Analysis of antioxidative phenolic compounds in artichoke
690 (*Cynara scolymus* L.). J Agr Food Chem 51:601-608
- 691 Wiklund A (1992) The genus *Cynara* L. (Asteraceae, Cardueae). Bot J Linn Soc 109:75-123
- 692 Yasukawa K, Matsubara H, Sano Y (2010) Inhibitory effect of the flowers of artichoke (*Cynara*
693 *cardunculus*) on TPA-induced inflammation and tumor promotion in two-stage carcinogenesis
694 in mouse skin. J Nat Med 64:388-391
- 695 Ye S, Dhillon S, Ke X et al. (2001) An efficient procedure for genotyping single nucleotide
696 polymorphisms. Nucleic Acids Res 29:e88-8

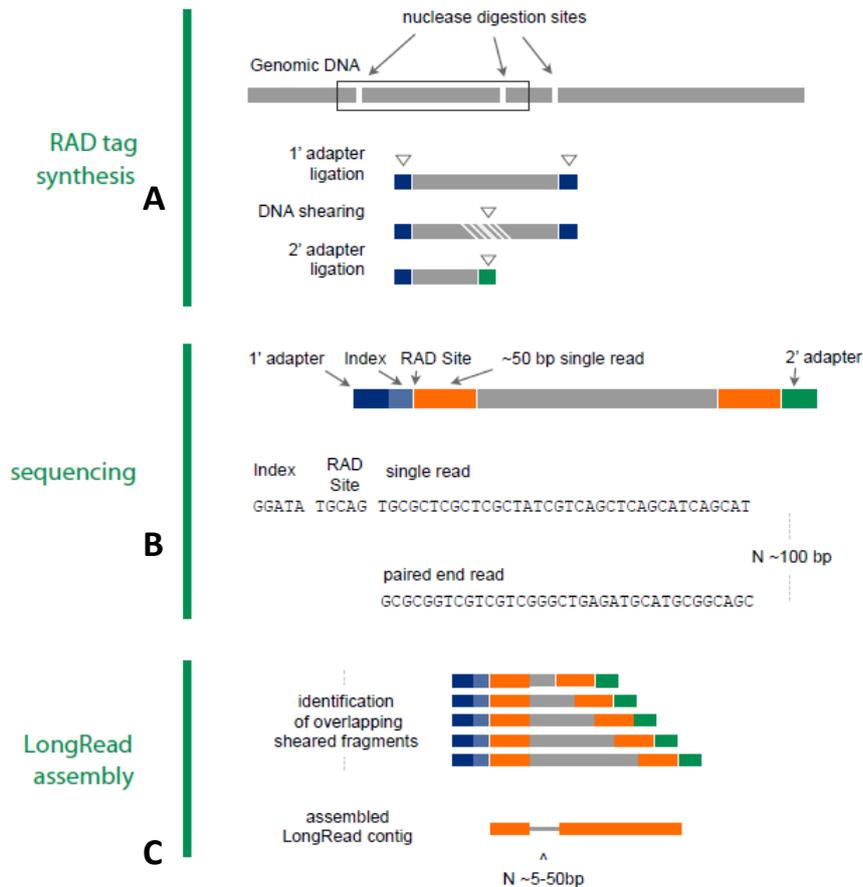
BOX 1

RADseq (Restriction-site Associated DNA sequencing)

An efficient approach for SNP discovery, is RAD “Restriction-site Associated DNA” (Miller et al. 2007), coupled with NGS technologies (Baird et al. 2008), which has been recently termed as RADseq (Davey et al. 2011). At least 20 papers have been recently published in both animals (snails, moths and salmon, sturgeon, butterflies, beetles and worms) and plants (ryegrass, oaks, lolium, eggplant and globe artichoke). A detail review is available at the wiki RAD-sequencing page (University of Edinburgh; <https://www.wiki.ed.ac.uk/display/RADSequencing>).

The strategy requires the enzymatic digestion of a genome with at least one restriction enzyme and the sequencing of the resulting fragments through an Illumina Genome Analyzer. The fragments from one sample are ligated to a modified Illumina adapter containing a unique identifying sequence (Molecular IDentifier, or MID). A list of the available primers can be found at the above-cited wiki RAD-sequencing section. The fragments from many samples (e.g. a mapping population) can consequently be pooled together and sequenced on a single lane. The resulting reads can be segregated using the MID present at the start of each read. By sequencing simultaneously all the individuals of a population of interest, and by comparing the tags, thousand of SNPs at different genetic loci can be identified in a single experiment.

The protocol is depicted in the figure reported below. **A)** Genomic DNA is digested with a restriction enzyme and a barcoded P1 adapter is ligated to the fragments. The P1 adapter contains a forward amplification primer site, an Illumina sequencing primer site, and a barcode for sample identification. Adapter-ligated fragments are pooled (if multiplexing), sheared, size-selected (e.g. 300-800 bp) and ligated to a second adapter (P2). The P2 adapter is a divergent “Y” adapter, containing the reverse complement of the P2 reverse-amplification primer site, preventing amplification of genomic fragments lacking a P1 adapter. **B)** The samples are analysed on an Illumina Genome Analyzer IIx following the paired ends (2x 54 bp, or more) genomic DNA sequencing protocol. The generated sequences are then sorted according to their multiplex identifier tag (barcode). **C)** The sequences are *de novo* assembled using a bioinformatics DNA assembler (e.g.: Velvet). Assembled LongRead® contigs can be generated by a set of algorithms developed at Floragenex Inc. (Oregon, USA).



GO ID	Term	N° of loci	EST-SSR loci
GO:0050896	response to stimulus	17	CyEM-008, -030, -42, -054, -057, -070, -072, -093, -120, -135, -145, -150, -152, -218, -229, -259, -266
GO:0009628	response to abiotic stress	13	CyEM -008, -030, -054, -070, -093, -120, -135, -145, -150, -152, -218, -229, -259
GO:0042221	response to chemical stimulus	4	CyEM -093, -218, -229, -266
GO:0006950	response to abiotic stress	15	CyEM -008, -030, -054, -057, -070, -072, -093, -120, -135, -145, -150, -152, -229, -259, -266
GO:0009266	response to temperature stress	5	CyEM -008, -054, -093, -145, -150
GO:0006970	response to osmotic stress	8	CyEM -030, -070, -093, -120, -135, -152, -229, -259
GO:0010033	response to organic substance	3	CyEM -093, -229, -266
GO:0009409	response to cold stress	5	CyEM -008, -054, -093, -145, -150
GO:0009651	response to salt stress	8	CyEM -030, -070, -093, -120, -135, -152, -229, -259

Table 1: CyEM (*Cynara* Expressed Microsatellites) markers with Gene Ontology annotation for stimuli response-related terms.

#	Genotype	<i>C. cardunculus</i> taxon	Sequencing results			Assembly results	
			Raw reads	Total (Mbp)	Mean length (bp)	Contigs	Mean length/N50 (bp)
1	‘Romanesco C3’	var. <i>scolymus</i>	0.43 M	184	421	37,622	834/723.8
2	‘Altilis 41’	var. <i>altilis</i>	0.61 M	246	402	40,130	761/699.9
3	‘Creta 4’	var. <i>sylvestris</i>	0.69 M	263	377	42,837	772/688.5
Total			1.74 M	693	392	38,726*	951/844.3*

Table 2. 454-derived sequencing and assembly. The output statistics were calculated following the removal of contaminating and adaptor sequences. Data are intended after quality filtering and sequence clipping. *Asterisks indicate results obtained by merging the three independent assemblies (see Figure 3).

#	Genotype	<i>C. cardunculus</i> taxon	Raw reads	First mates (Mbp)	Paired mates (Mbp)
4	'Romanesco Zorzi'	var. <i>scolymus</i>	6.6M	458	408
5	'Violetto di Chioggia'	var. <i>scolymus</i>	6.6M	470	420
6	'Violetto Pugliese'	var. <i>scolymus</i>	5.2M	367	331
7	'Spinoso Sardo'	var. <i>scolymus</i>	6.7M	474	424
8	'Imperial Star'	var. <i>scolymus</i>	6.4M	459	415
9	'Blanco de Peralta'	var. <i>atilis</i>	4.8M	340	305
10	'Gobbo di Nizza'	var. <i>atilis</i>	5.6M	380	341
11	'Sylvestris_LOT23'	var. <i>sylvestris</i>	4.6M	322	287
Total			46.5M	3,271	2,931

Table 3. GAIIX (Illumina)-derived sequencing. A total of 46.5 M raw reads were generated in two GAIIX lanes and 6.7 Gbp were retained after removing adaptor and contaminating sequences. The windowed quality clipping routine produced a final dataset of 6.2 Gbp. A higher number of bases were obtained for single ends, because 84 sequencing cycles were used instead of the 76 used for the paired ends.

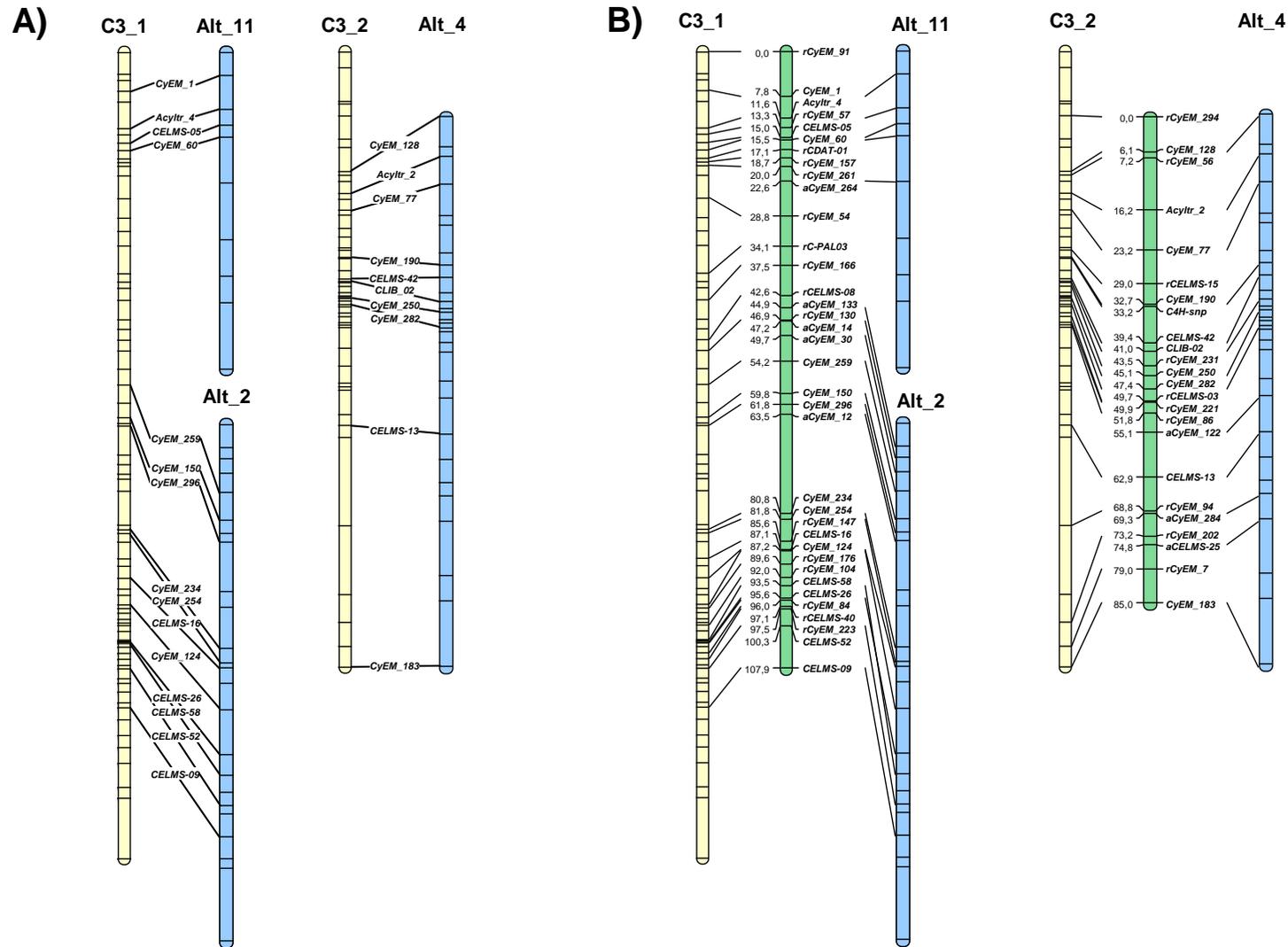


Figure 1: Examples of alignment and consensus LG construction. Alignment of the ‘Romanesco C3’ (yellow) and the ‘Altis 41’ (blue) LGs based on common markers (A). SSR-based consensus LGs (green) construction (B). ‘r-’ and ‘a-’ indicate markers segregating only in, respectively, ‘Romanesco C3’ and ‘Altis 41’. Marker nomenclature is the one reported in Portis et al. (2009) and Scaglione et al. (2009).

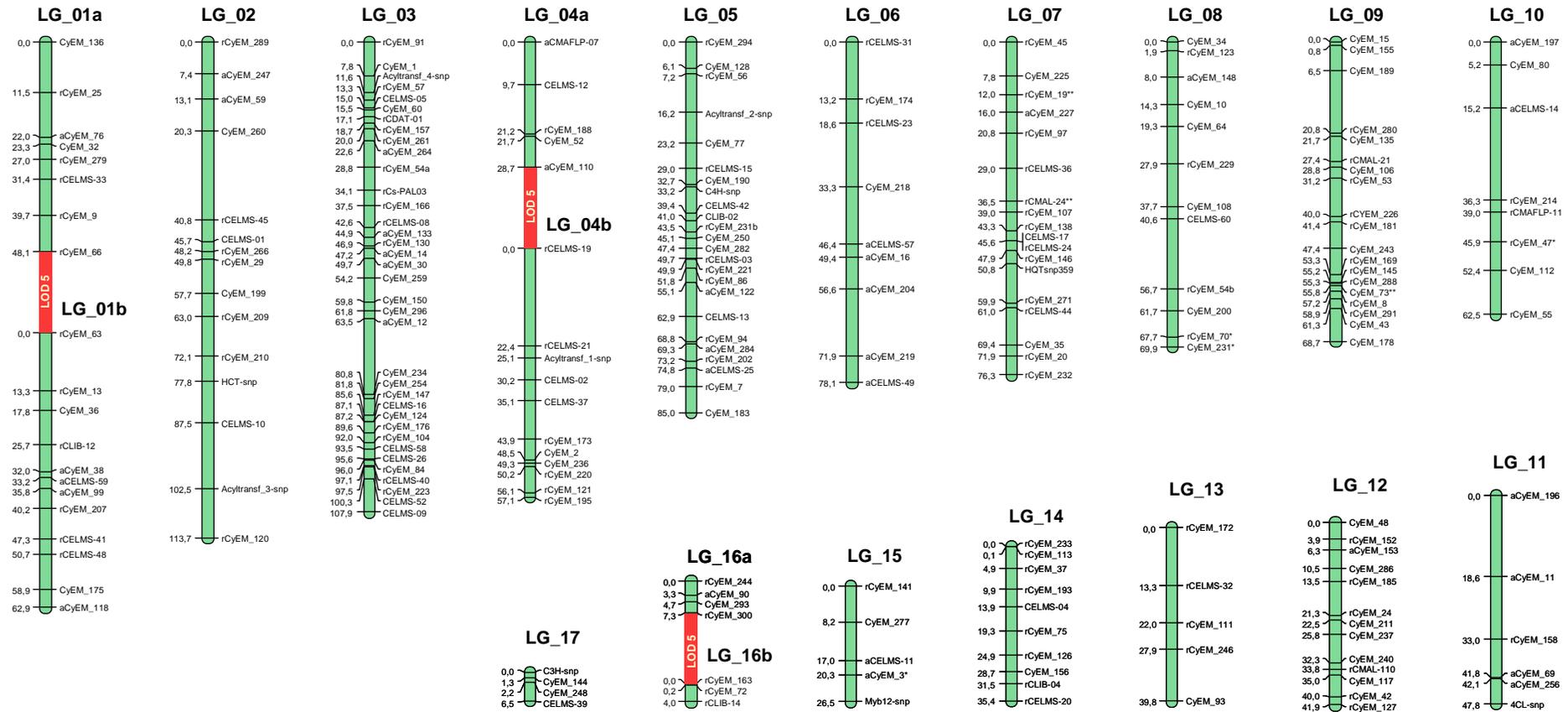


Figure 2: The SSR-based consensus map of *C. cardunculus*. Marker names are shown to the right of each LG, with map distances (in cM) to the left. 'r-' and 'a-' indicate markers segregating only in, respectively, 'Romanesco C3' and 'Altilis 41'. Segments shaded in red indicate where a pair of LGs has merged as a result of reducing the stringency to LOD 5. Marker nomenclature is the one reported in Portis et al. (2009) and Scaglione et al. (2009).

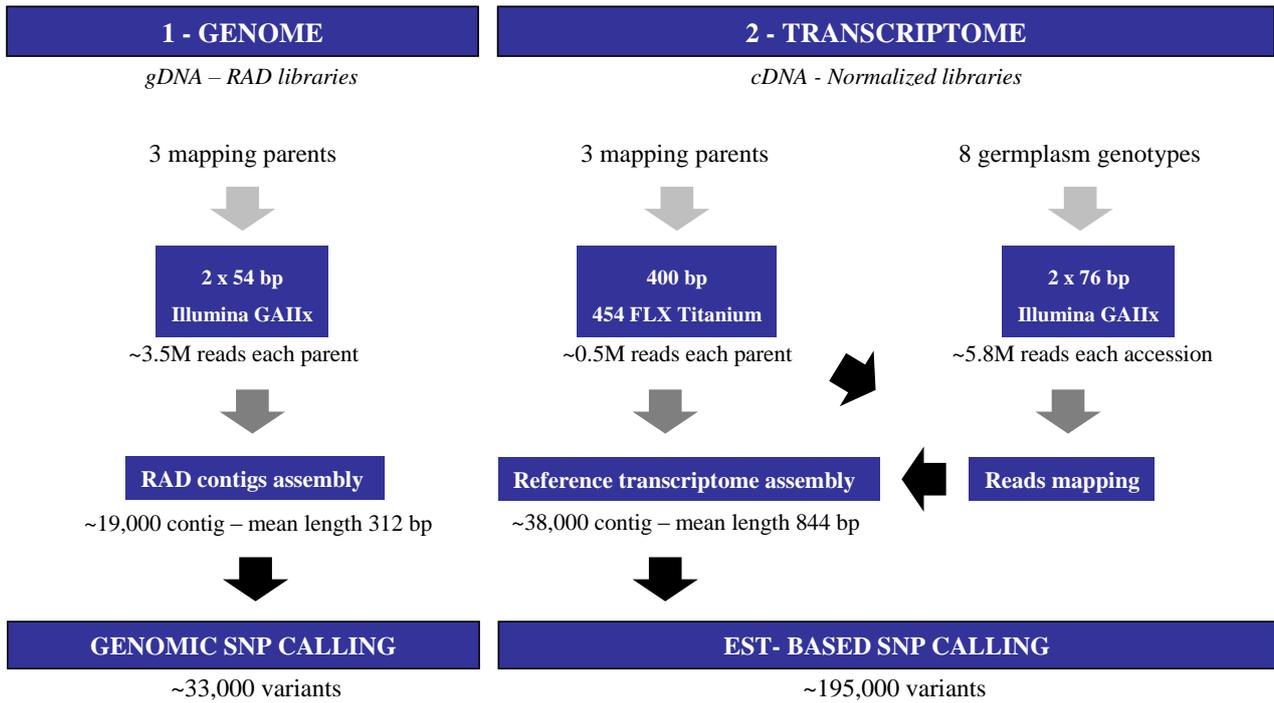


Figure 3. SNP mining workflow in *Cynara cardunculus*.

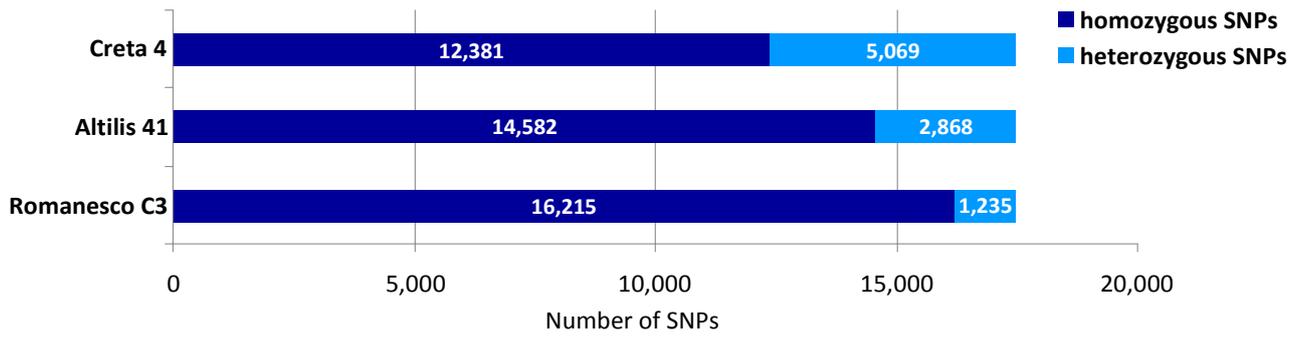


Figure 4: Proportion of heterozygous SNPs across the three mapping parents.

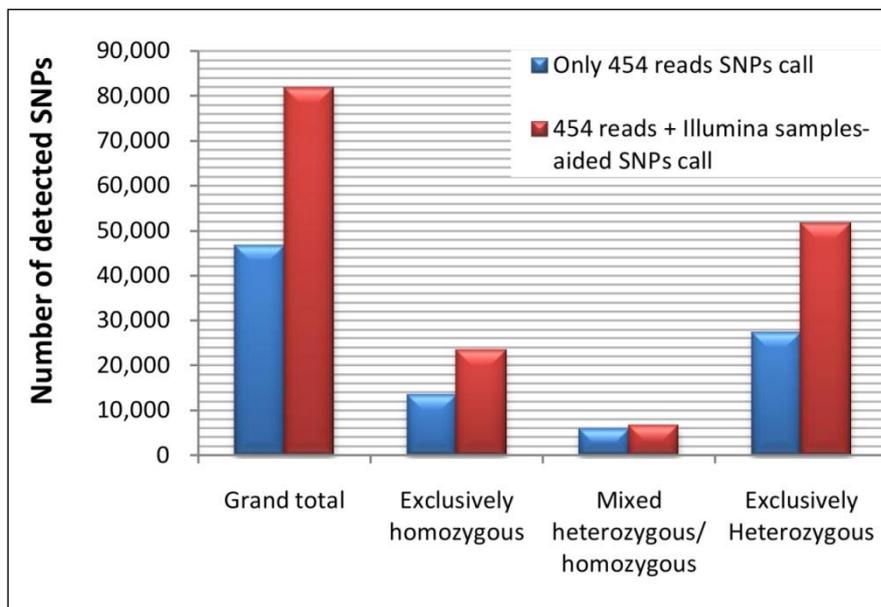


Figure 6: Combined calling of SNPs. The number of calls based solely on the 454-derived reads is shown in blue, and the combined SNP discovery based on both the 454- and the Illumina-based sequence in red. “Exclusively homozygous” and “exclusively heterozygous” refer to allelic variants present in only one of the three 454-sequenced libraries.

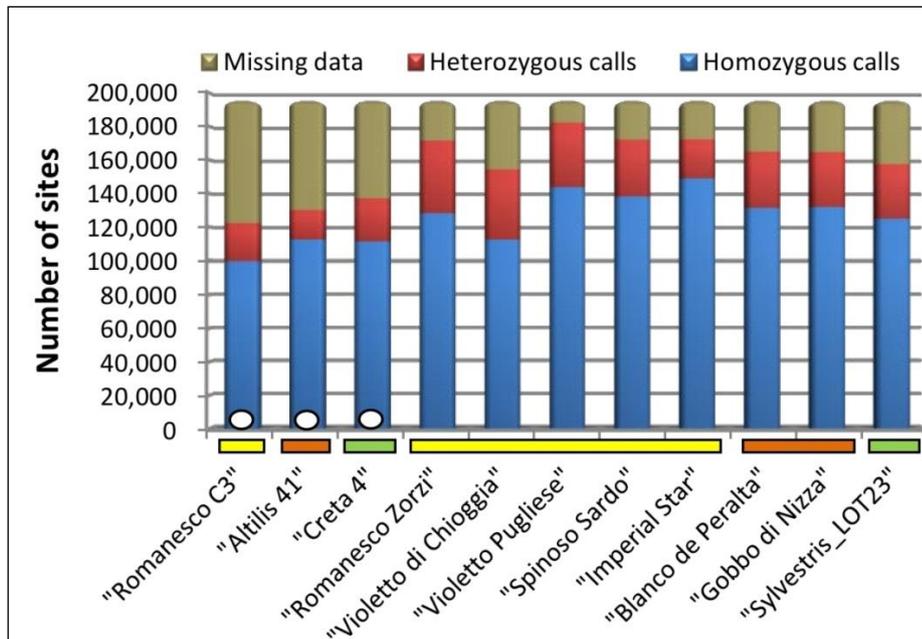


Figure 7: The allelic state at SNP loci. Bars indicate the total number of SNP loci in the homozygous or heterozygous state (or missing) for each accession. Each bar's colour identifies the *C. cardunculus* taxa (green = *sylvestris*, orange = *altilis*, yellow = *scolymus*). White dots identify the three accessions sequenced using 454 technology.