Provisional chapter

Toward a First High-quality Genome Draft for Marker-assisted Breeding in Leaf Chicory, Radicchio (*Cichorium intybus* L.)

Giulio Galla, Andrea Ghedina, Silvano C. Tiozzo and Gianni Barcaccia

Additional information is available at the end of the chapter

Abstract

Radicchio (*Cichorium intybus* subsp. *intybus* var. *foliosum* L.) is one of the most important leaf chicories, used mainly as a component for fresh salads. Recently, we sequenced and annotated the first draft of the leaf chicory genome, as we believe it will have an extraordinary impact from both scientific and economic points of view. Indeed, the availability of the first genome sequence for this plant species will provide a powerful tool to be exploited in the identification of markers associated with or genes responsible for relevant agronomic traits, influencing crop productivity and product quality. The plant material used for the sequencing of the leaf chicory genome belongs to the Radicchio of the Chioggia type. Genomic DNA was used for library preparation with the TruSeq DNA Sample Preparation chemistry (Illumina). Sequencing reactions were performed with the Illumina platforms HiSeq and MySeq, and sequence reads were then assembled and annotated. We are confident that our efforts will extend the current knowledge of the genome organization and gene composition of leaf chicory, which is crucial for developing new tools and diagnostic markers useful for our breeding strategies in Radicchio.

Keywords: Genome draft, marker-assisted breeding, gene prediction, SSR markers, SNP calling

1. Introduction

The common Italian name of Radicchio was adopted in recent years by all the most internationally used languages and indicates a highly differentiated group of chicories, with red or variegated leaves. Radicchio (*Cichorium intybus* subsp. *intybus* var. *foliosum* L.) is currently one of the most important leaf chicories, used mainly as a component for fresh salads but also very

often cooked and prepared differently according to local traditions and alimentary habits [1]. This plant species belongs to the Asteraceae family and includes several cultivar groups whose commercial food products are the leaves, namely Witloof, Pain de sucre, and Catalogne, as well as several types of Radicchio.

From the reproductive point of view, Radicchio is prevalently allogamous, due to an efficient sporophytic self-incompatibility system, proterandry and gametophytic competition favoring allo-pollen grains and tubes [1]. Probably known by the Egyptians and used as food and/or medicinal plants by the ancient Greeks and Romans, this species gradually underwent a process of naturalization and domestication in Europe during the past few centuries. This plant has become part of both natural and agricultural environments of Italy. Currently, among the different biotypes of leaf chicories, the so-called Radicchio of Chioggia, native to and very extensively grown in northeastern Italy, is the Radicchio cultivar acquiring more and more commercial interest worldwide. In Italy, the Radicchio of Chioggia is cultivated on a total area of approximately 16–18,000 ha, half of which is in the Veneto region, with a total production of approximately 270,000 tons (more than 60% obtained using professional seeds), reaching an overall turnover of approximately € 10,000,000 per year.

Grown plant materials are usually represented by landraces or their directly derived synthetics that are known to possess a high variation and adaptation to the natural and anthropological environment where they originated from and are still cultivated. These populations are characterized by high-quality traits and have been maintained or even improved over the years by local farmers through phenotypical selection according to their own criteria and more recently by seed companies through genotypical selection following intercross or polycross schemes combined with progeny tests to obtain populations showing superior DUS scores for both agronomic and commercial traits. The breeding programs currently underway by local firms and regional institutions exploit the best landraces and aim to isolate individuals amenable for use as parents for the constitution of narrow genetic base synthetic varieties and/or to select inbred lines suitable for the production of heterotic F1 hybrids [2]. In recent years, phenotypic evaluation trials are increasingly assisted by genotypic selection procedures through the use of molecular markers scattered throughout the genome. In fact, marker-assisted breeding allows the identification of the parental individuals or the inbred lines showing the best general or specific combining ability in order to breed synthetics and hybrids, respectively.

Radicchio, like the other leaf chicories, is diploid ($2n=2x=18$) and is characterized by an estimated haploid genome size of approximately 1.3 Gb. In recent years, three distinct saturated molecular linkage maps were constructed for leaf chicories, covering approximately 1,200 cM [3-5]. Its linkage groups were mainly based on neutral SSR markers, but many EST-derived SNP markers were also mapped. A method for genotyping elite breeding stocks of Radicchio, both local and modern varieties, assaying mapped SSR marker loci possibly linked to EST-rich regions and scoring PIC>0.5, was recently developed using multiplex PCRs [6]. Here, we are dealing with a research and development project aimed at sequencing and annotating the first draft of the leaf chicory genome as we believe it will have an extraordinary impact from both scientific and economic points of view. Indeed, the availability of the first

genome sequence for this plant species will provide a powerful tool to be exploited in the identification of markers associated with or genes responsible for relevant agronomic traits, influencing crop productivity and product quality. As an example, data and knowhow produced in this research project will be useful for detailed studies of the genetic control of male-sterility and self-incompatibility in this species.

The plant material that we used for the sequencing of the leaf chicory genome belongs to the Radicchio of Chioggia type, specifically to the male fertile inbred line named SEG111. This type was chosen as the most suitable accession based on the following criteria: i) the commercial relevance of the variety of origin; ii) the availability of clonal materials; iii) robust phenotypic and genotypic characterization; iv) a high degree of homozygosity (80%); and v) high breeding value as pollen parent of F1 hybrids. Sequencing reactions of the genomic DNA library were performed with Illumina HiSeq and MySeq platforms to combine the high number of reads originated by the former with the longer sequences produced by the latter. Here, we report original data from the bioinformatic assembly of the first genome draft of Radicchio, along with the most relevant findings that emerged from an extensive *de novo* gene prediction and *in silico* functional annotation of more than 18,000 unigenes. Analyses were performed according to established computational biology protocols by taking advantage of the publically available reference transcriptome data for *Cichorium intybus* [7]. The main preliminary findings on the genome organization and gene composition of Radicchio are presented, and the potentials of newly annotated expressed sequences and diagnostic microsatellite markers in breeding programs are critically discussed.

2. Materials and methods

2.1. Plant materials

Plant materials used for the sequencing belong to a variety of commercial relevance of the Radicchio of Chioggia type. The clone chosen derives from the inbred line SEG111 and shows a degree of homozygosity equal to 80% [6]. In particular, this clone was obtained by several cycles of selfing from plants yearly selected on the basis of a robust phenotypic and genotypic characterization, being also characterized by high-quality agronomic traits on farm and the ability to be easily cloned in vitro.

2.2. DNA isolation and sequencing

DNA was isolated from 150 mg of fresh leaf tissue using a CTAB-based protocol [8]. The eventual contamination of RNA was avoided with an RNase A (Sigma-Aldrich) treatment. DNA samples were eluted in 80–100 μL of 0.1× TE buffer (100 mM Tris-HCl 1, 0.1 mM EDTA, pH=8). The integrity of the extracted DNA samples was estimated through electrophoresis in 0.8% agarose/1× TAE gels containing 1× SYBR Safe DNA Gel Stain (Life Technologies, USA). The purity and quantity of the DNA extracts were assessed with a NanoDrop spectrophotometer (Thermo Scientific, USA). Then, 1 μg of high-quality DNA

was used for library preparation with the TruSeq DNA Sample Preparation chemistry (Illumina). Sequencing reactions were performed with the Illumina platforms: HiSeq (1 lane, 2 × 100 bp) and MySeq (1 lane, 2 × 300 bp).

2.3. *De novo* assembly and annotation

All high-quality reads generated from the two sequencing reactions were assembled in a single reference genome. Assemblies were attempted with three pieces of software: i) Velvet [9]; ii) SPAdes [10]; and iii) CLC Genomics Workbench 6.5 (Qiagen). The average coverage was estimated for the run HiSeq by calculating the frequency distribution of 25-mers [11].

To annotate all assembled contigs, a BLASTX-based approach was used to compare the *C. intybus* sequences to a subset of the NR protein collection that was made by focusing on the clade pentapetalae [12]. Moreover, the GI identifiers of the best BLASTX hits, having E-value ≤1.0E-15 and similarity ≥70%, were mapped to the UniprotKB protein database [13] to extract Gene Ontology annotations [14] and KEGG terms [15] for functional annotations. Further enrichment of enzyme annotations was made with the BLAST2GO software v1.3.3 using the function "direct GO to Enzyme annotation". The BLAST2GO software v1.3.3 [16, 17] was used to reduce the complexity of the data and perform basic statistics on ontological annotations, as reported by Galla *et al.* [18].

SSRs were detected among the 522.301 contigs via MISA [19]. The parameters were adjusted to identify perfect and complex mono-, di-, tri-, tetra-, penta-, and hexanucleotide motifs with a minimum of 49, 13, 9, 8, 8, and 8 repeats, respectively. Repeated elements were detected with a BLASTN-based approach using a PGSB Repeat Element Database in all blast searches [20]. The parameters set for the identification of Transposable Elements (TEs) were: reward 1, penalty 1, gap_open 2, gap_extend 2, word_size 9, dust no. An E-value cutoff of 1.0E-9 was adopted to filter the BLAST results.

Two public *C. intybus* transcriptomes CHI-2418 and CHI-Witloof originally developed from plant seedlings [7] corresponding to a wild accession of leaf chicory and a cultivated variety of witloof, respectively, were mapped to the reference genome using the CLC Genomics Workbench V7.02 (Qiagen). Mappings were performed with default mapping parameters, including mismatch cost: 2; insertion cost: 3; deletion cost: 3; length fraction: 0.5; and similarity fraction: 0.8. Non-specific matches were ignored and not included in the annotation tracks. For nucleotide variant analysis, the appropriate reference masking options were used to map transcriptome reads selectively over the sequences annotated as CDS or TEs. The variant detection analysis was done by using the Basic Variant Detection tool of the CLC Genomics Workbench V7.02 (Qiagen) with default parameters. As general filters, positions with coverage above 100,000 were not considered. Base quality filters were turned on and set to default parameters. All variants included in homopolymer regions with minimum length of 3nt, and with frequency below 0.8 were also removed from the dataset. As coverage and count filters, all variants with a minimum count lower than 20 were discarded.

3. Results

3.1. Genome assembly statistics

To obtain the first genome draft of leaf chicory, a single genomic library produced from the inbred line SEG111 was sequenced using the Illumina MySeq and HiSeq platforms. Here, we report the genome assembly results derived from the CLC Genomic Workbench assembly output. Figure 1 describes the frequency distribution of 25-mers in the HiSeq data.

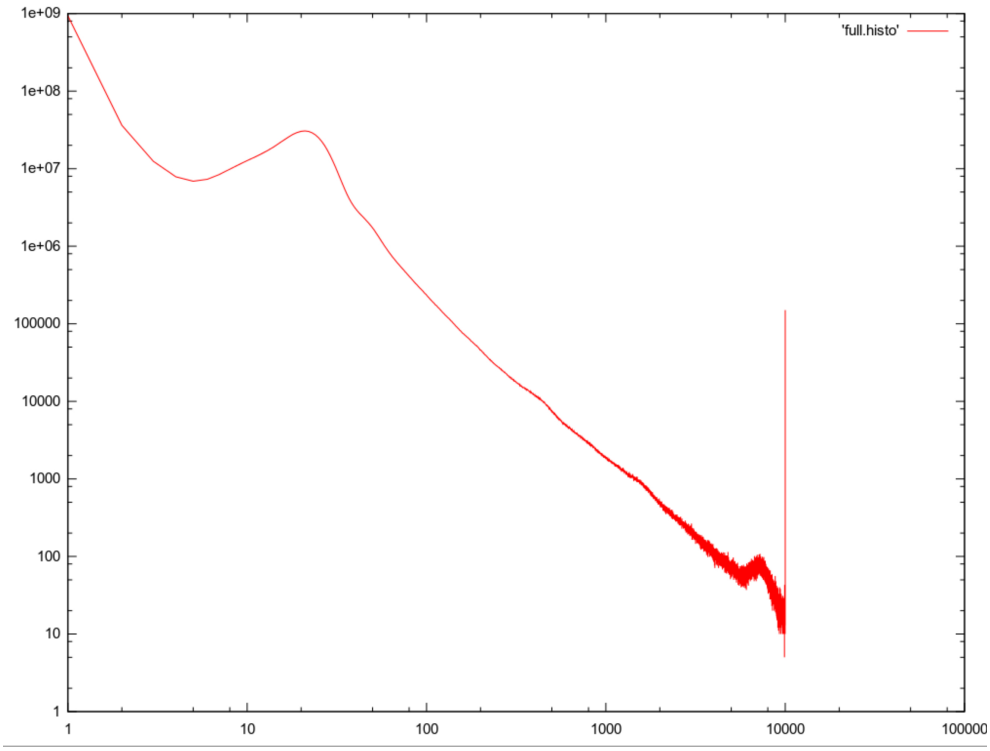

Figure 1. Frequency distribution of 25-mers in the HiSeq data (logarithmic scale for both axes)

The data shown suggest that the average coverage in the HiSeq run is approximately 21×. Additionally, the curve indicates that a certain number of sequences are present with a relatively high frequency within the genome. This might indicate that repeated elements are relatively abundant within the genome. As a consequence, the estimated size of the assembled genome draft is 760 Mb.

We obtained 58,392,530 and 389,385,400 raw reads through the MySeq and HiSeq platforms, respectively. The *de novo* assembly of the two datasets in a unique reference genome draft assembled 724,009,424 nucleotides into 522,301 contigs (Table 1). The maximum contig length

was equal to 379,698 bp, whereas the minimum contig length was set to 200 bp, with an average contig length of 1,386 bp. Overall statistics are summarized in Table 1.

Total number of contigs	522,301
Total No. of assembled nucleotides (nt)	724,009,424
GC percentage	34.8%
Average contig length (bp)	1,386
Minimum contig length (bp)	200
Maximum contig length (bp)	379,698
N75	1,051
N50	3,131

Table 1. Summary statistics of the sequence assembly generated from *Cichorium intybus*.

The length distribution of the contig size, expressed in base pairs, is reported in Figure 2.

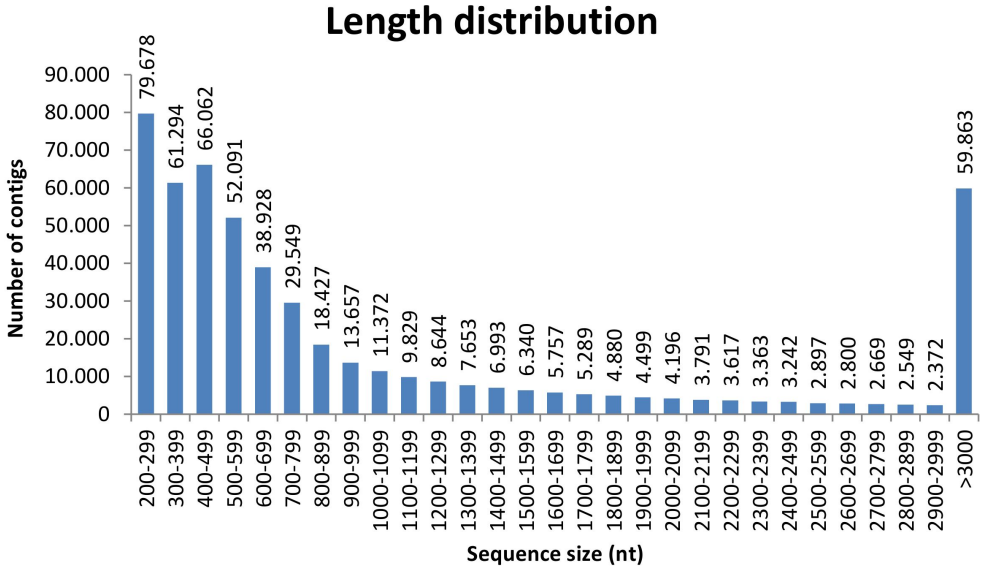

Figure 2. Distribution of length of contigs for leaf chicory

As much as 68.9% of the recovered sequences are contained within a length spanning from 200 nt to 999 nt. The interval length ranging between 1,000 nt and 2,999 nt is represented by 19.7% of the assembled contigs, whereas the proportion of contigs whose length is higher or equal to 3,000 nt corresponds to 11.5%.

We searched the genome sequence assembly for TEs and estimated their abundance using a BLASTN strategy. The proportion of base pairs annotated as TEs out of the total amount of assembled nucleotides was equal to 6.3% (Table 2).

Key	Classification	Number	Abundance (%)	Length (bp)	Percentage over the assembled genome
02.01	Class I retroelement	273	0.19	85,241	0.012%
02.01.01	LTR Retrotransposon	82,260	56.55	19,658,874	2.715%
02.01.01.05	Ty1/copia	35,802	24.61	17,519,102	2.420%
02.01.01.10	Ty3/gypsy	23,651	16.26	7,121,605	0.984%
02.01.02	non-LTR Retrotransposon	354	0.24	106,259	0.015%
02.05	Class II: DNA Transposon	1,976	1.36	713,119	0.098%
02	Unclassified mobile element	861	0.59	199,301	0.028%
10 / 90 / 99	High Copy Number Genes and additional attributes	283	0.19	51,577	0.007%
Total		145,462	100.0	45,455,078	6.278%

Table 2. Classification statistics of transposable elements (TEs) in Radicchio genome draft assembly.

The retroelements were the most abundant elements (>97% of the total). Within the major class of retroelements, Long Terminal Repeat (LTR) retrotransposons proved to be the dominant class (56.55%) in the leaf chicory genome. Moreover, the Copia-type (24.61%) and the Gypsy-type (16.26%) appeared to be the most abundant LTR retrotransposons. A total of 273 (0.2%) elements were annotated as retroelements, but they lacked the assignation to a specific class based on sequence similarity and conservation. Non-LTR retrotransposons were detected to a very low extent (0.24%). Less than 2% of the total repeat elements were annotated as DNA transposons.

3.2. Discovery of SSR loci

Overall, we identified 66,785 SSR containing regions. As many as 52,186 and 11,501 sequences proved to contain one or more microsatellites, respectively. These numbers included 1,226 mononucleotide SSR motifs (which were no longer taken into account for further computations). We found a total number of di- or multinucleotide SSR motifs equaling 65,559.

The most common SSR elements were those showing a dinucleotide motif (89.0%), followed by trinucleotide (7.1%) and tetranucleotide (3.0%) ones. Microsatellites revealing a pentanucleotide and hexanucleotide motif were less than 1.0% of the total. Overall data are summarized in Table 3.

Type of motif	Range of repeat numbers				Total No.	Percentage (%)
	8-12	13-17	18-22	>22		
Di-nucleotide	0	8,333	7,100	42,913	58,346	89.0
Tri-nucleotide	1,822	1,769	762	321	4,674	7.1
Tetra-nucleotide	1,114	475	205	202	1,996	3.0
Penta-nucleotide	69	23	0	2	94	0.1
Hexa-nucleotide	359	80	8	2	449	0.7
Total	3,364	10,680	8,075	43,440		
Percentage (%)	5.1	16.3	12.3	66.3		

Table 3. Number of SSRs detected in the Radicchio genome draft assembly. For each type of motif, the number of SSRs identified in the range of repeated numbers is reported. Albeit present in the genome, mono-nucleotide SSRs were not considered in this analysis.

3.3. Functional annotation of contig sequences

Functional annotation of the assembled contigs was performed with a BLASTX approach, according to which all contig sequences were used to query different public protein databases (Table 4).

Public database	No. of Hits (gene models)	No. of *C. intybus* contigs
NR	38,782	80,862
Arabidopsis	16,689	50,417
GO	14,073	45,381
KEGG	4,512	22,273

Table 4. Summary statistics of functional annotations for leaf chicory genome sequences in public protein databases. As for the NR database, only the protein sequences from the clade pentapetalae of eudicots were considered. The Arabidopsis proteome used in all BLAST analysis was TAIR10.

The database enclosing all public protein sequences belonging to the pentapetalae clade of the eudicots, which includes the sub-clades of rosids and asterids to which leaf chicory belongs, provided a total of 38,782 hits. The proteome of *Arabidopsis thaliana* alone scored 16,689 hits when an E-value cutoff of 1.0E-15 was applied for the screening of the most reliable BLASTX hits.

Two public *C. intybus* transcriptomes originally developed from plant seedlings and provided by UC DAVIS, the Compositae Genome Project (CHI-2418 and CHI-Witloof) [7] were mapped to the reference genome using the appropriate mapping function of the CLC Genomics Workbench.

By doing so, we were able to map 76.5% and 78.0% of the sequences, respectively. Data derived from the mapping of two *C. intybus* transcriptomes were used to integrate the annotation of the assembled contigs. BLAST and mapping data integration increased the BLAST-based annotation with an additional set of 1,995 contigs.

Arabidopsis matches were used to retrieve both GO and KEGG annotations from public databases. We could finally assign one or multiple GO terms to 45,381 leaf chicory genome contigs. The analysis performed against the GO illustrate 14,073 genes annotated with terms belonging to one or multiple vocabularies. Of these, 24,634 contigs were annotated for their putative biological process, 39,118 contigs were related to a molecular function, and 37,561 contigs were associated to a specific cellular component. Figure 3 shows the fine distribution of the 14,073 hits caught by our Radicchio contigs from the TAIR database according to the aforementioned three GO categories.

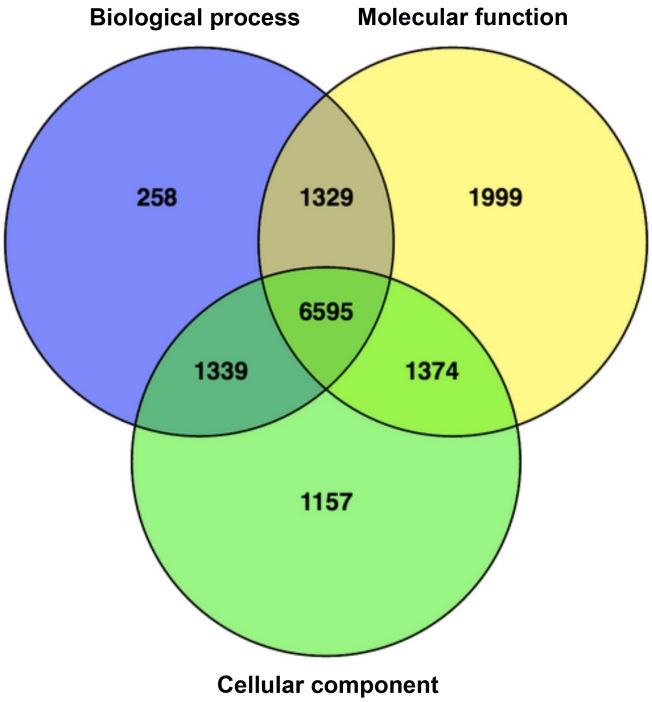

Figure 3. Venn diagram showing the fine distribution according to GO terms of the 14,073 *A. thaliana* hits matching our leaf chicory contigs

Among all the terms underlined by the GO vocabulary for the biological process, our investigations were focused on terms related to the response to biotic and abiotic stresses (Figure 4), hormonal responses (Figure 5), and flower and seed development (Figure 6). Of the 15 most

interesting processes for molecular breeding in leaf chicory, 7 and 8 were linked to biotic and abiotic stresses, respectively (see Figure 4). The ontological terms were assigned to 2,388 and 3,844 genome contigs, respectively.

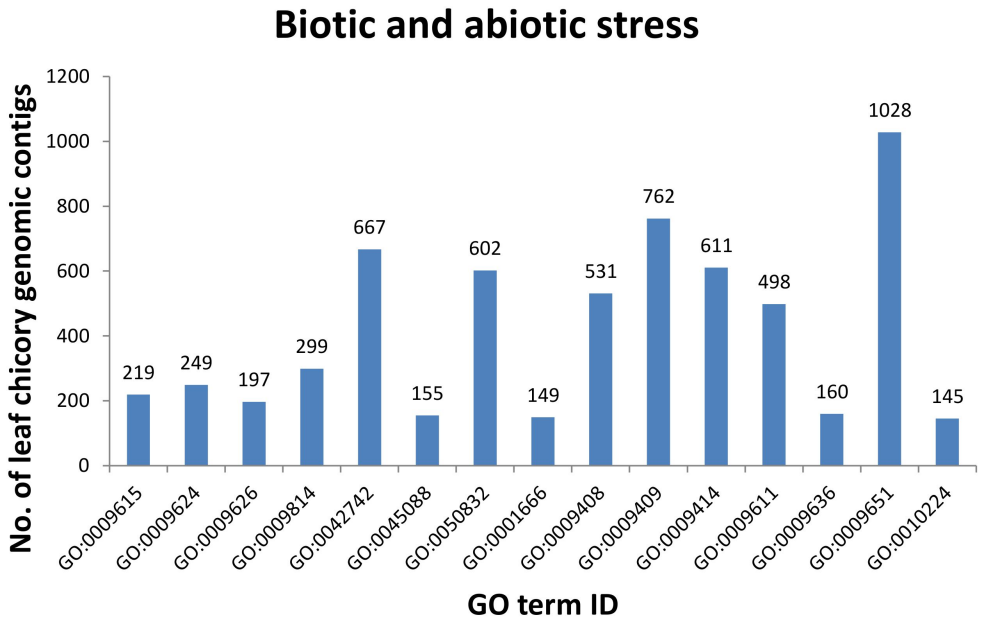

Figure 4. Number of *C. intybus* genomic contigs for response to biotic (the first 7) and abiotic (the last 8) stress

The computational analysis for the identification of SSR elements within these contigs unveiled 495 motifs linked to biotic stresses and 841 motifs associated with abiotic stresses. Among the biotic stresses, the most abundant gene ontology (GO) term was GO:0042742, which corresponds to the "defense response to bacterium" and shows a match with 667 genome contigs containing 135 microsatellites. Concerning the abiotic stresses, the GO term assigned with the higher frequency was GO:0009651, which accounts for processes related to "response to salt stress" and matches 1,028 genome contigs containing 249 microsatellites.

Data of hormonal responses and processes of flower and seed development are reported in Figures 5 and 6. The analysis for hormonal responses noted nine different GO terms, for a total of 3,344 genome contigs, and 833 SSR elements linked to these sequences and terms. In particular, the term "response to jasmonic acid stimulus" (GO:0009753) was the most represented, with 478 matches with different genome contigs, including 118 SSR motifs (Figure 5).

Results of the GO term annotation of genome contigs according to the flower and seed developmental processes are reported in Figure 6.

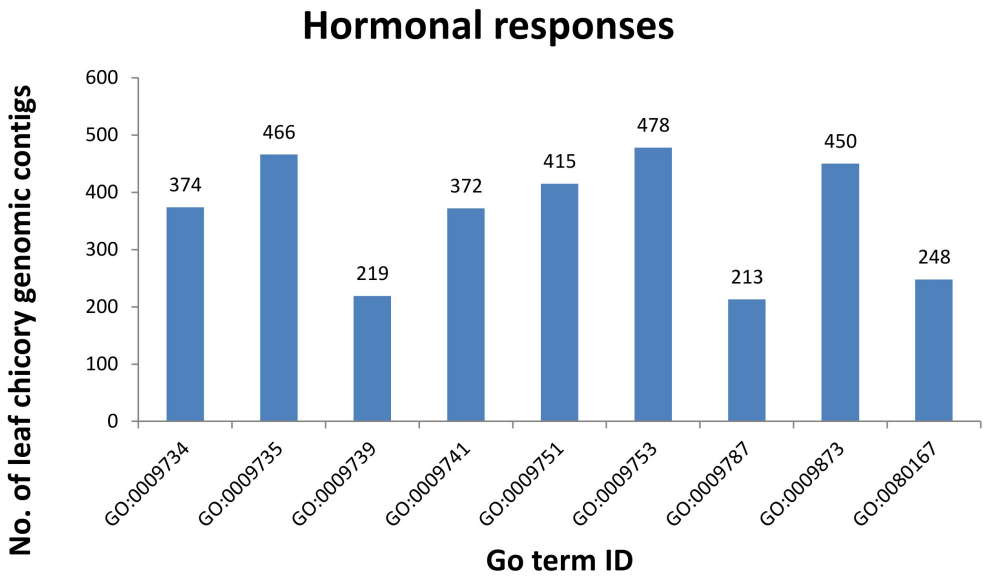

Figure 5. Number of *C. intybus* genomic contigs for hormonal response

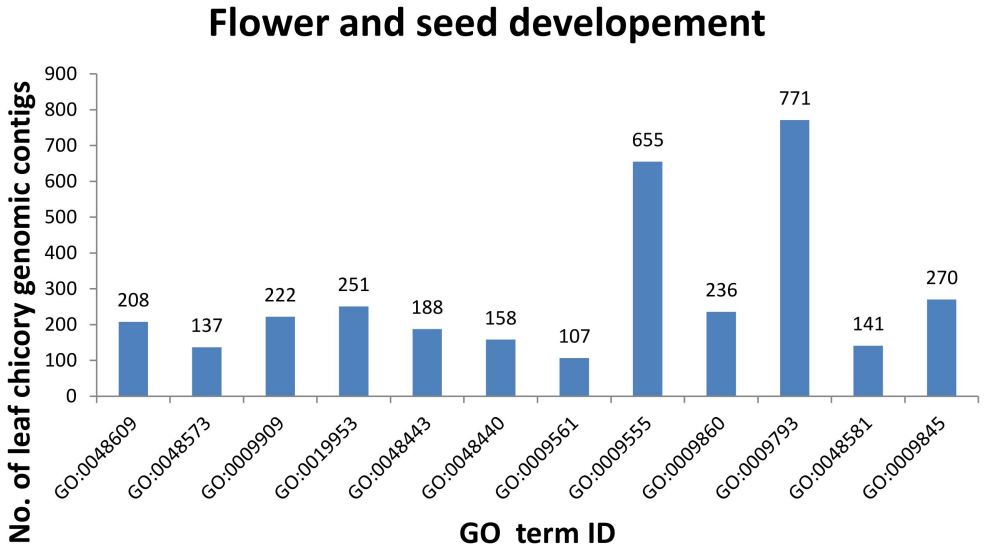

Figure 6. Number of *C. intybus* genomic contigs for flower and seed (only the last 3) development

The flower development process was embraced by selecting nine ontological terms, whereas three terms were assigned to seed development and seed germination. A total of 2,162 contigs

were annotated with GO terms related to flower development; 496 of these were also annotated for the presence of one or multiple SSRs. In particular, the term "pollen development" (GO: 0009555) was the most abundant, with 655 contigs containing 153 SSR motifs.

As far as the seed development process is concerned, we annotated 1,182 contigs linked to this GO term, 273 of which co-localized with one or multiple SSRs. Among these, the most abundant ontological term was "embryo development ending in seed dormancy" (GO: 0009793) as it is assigned to 771 contigs, co-localizing with 171 SSR elements.

Using the Kyoto Encyclopaedia of Genes and Genomes database (http://www.genome.jp/kegg/), a total of 22,273 contigs enabled the mapping of 795 enzymes to 157 metabolic pathways. Among the metabolic pathways with the highest number of mapped reads, we found fructose and mannose metabolism (418 gene models matched), phenylpropanoid biosynthesis (415 gene models matched) and tryptophan metabolism (380 gene models matched). The biosynthetic pathway of flavonoid biosynthesis, described in map:00941, is relevant as the biosynthesis of flavonoid is directly connected to the synthesis of anthocyanin (Figure 7), whose accumulation contributes to the pigmentation of leaf chicories. This map includes 236 gene models that were assigned to 14 unique enzymes, including CHS (CHALCONE SYNTHASE), CHI (CHALCONE ISOMERASE), and ANS (ANTHOCYANIDIN SYNTHASE), among others.

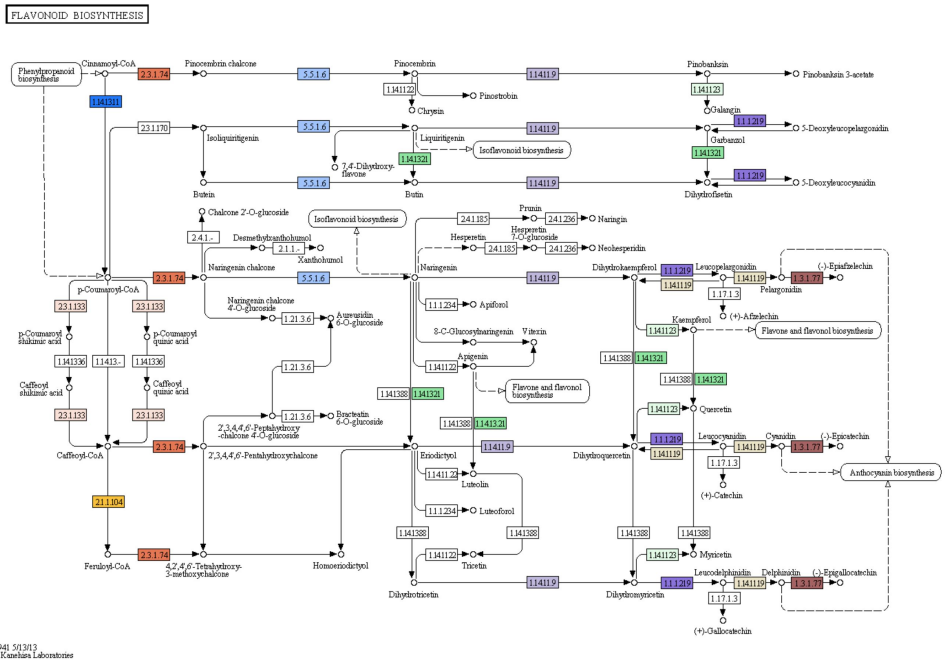

Figure 7. KEGG pathway for flavonoid biosynthesis (Map:00941)

KEGG data related to a number of selected metabolic pathways were exploited to find SSR regions potentially associated with highly valuable phenotypes in this plant species. The number of SSRs putatively linked to the most interesting phenotypic traits with breeding values in leaf chicory is displayed in Table 5.

KEGG map ID	Metabolic pathway	Characteristic	No. of SSRs
map00909	Sesquiterpenoid and triterpenoid biosynthesis	Bitter taste	107
map00053	Ascorbate and aldarate metabolism	Vitamin C content	172
map00940	Phenylpropanoid biosynthesis	Leaf color	281
map00941	Flavonoid biosynthesis	Leaf color	173
map00942	Anthocyanin biosynthesis	Leaf color	180
map00943	Isoflavonoid biosynthesis	Leaf color	5
map00944	Flavone and flavonol biosynthesis	Leaf color	128
map00040	Pentose and glucuronate interconversions	Response to cold	96
map00051	Fructose and mannose metabolism	Response to cold	259
map00052	Galactose metabolism	Response to cold	31
map00061	Fatty acid biosynthesis	Response to cold	39
map00260	Glycine, serine and threonine metabolism	Response to cold	60
map00290	Valine, leucine and isoleucine biosynthesis	Response to cold	13
map00330	Arginine and proline metabolism	Response to cold	55
map00410	beta-Alanine metabolism	Response to cold	16
map00480	Glutathione metabolism	Response to cold	48
map00500	Starch and sucrosa metabolism	Response to cold	164
map00561	Glycerolipid metabolism	Response to cold	159
map00564	Glycerophospholipid metabolism	Response to cold	124
map00592	alpha-Linolenic acid metabolism	Response to cold	66
map00710	Calvin cycle	Response to cold	28
map00780	Biotin metabolism	Response to cold	18
map00960	Tropane, piperidine and pyridine alkaloid biosynthesis	Response to cold	97

Table 5. Number of SSRs located in contig sequences annotated for the presence of proteins with known enzymatic activity in relevant metabolic pathways for the breeding of leaf chicory.

Considering the overall grouping of selected metabolic pathways, we identified many microsatellite sequences putatively linked to important traits, according to their potential effect on plant characteristics. For instance, 107 SSRs were linked to bitter taste, 172 SSRs were associated with vitamin C biosynthesis and metabolism, and 767 SSRs located in sequence contigs encoding enzymes of the flavonoid and anthocyanin biosynthetic pathways, thus potentially associated with the leaf color. The most represented characteristic is the response to cold. For this trait, we analyzed 16 different metabolic pathways that altogether led to the selection of 1,273 microsatellites potentially associated with one or multiple genes actively involved in the plant response to cold eventually, but not exclusively, through the accumulation of sugar.

We also performed the calling of nucleotide variants. Stringent quality criteria were used for discriminating sequence variations from sequencing errors and mutations introduced during cDNA synthesis. Only sequence variations with mapping quality scores over the established thresholds were annotated, leading to the identification of 123,943 and 121,086 variants that were present only in the leaf chicory transcriptome CHI-2418 (wild type) or the Witloof transcriptome CHI-Witloof (cultivated type), respectively. A total of 119,729 variants were shared by both *C. intybus* transcriptomes. The average number of variants per contig ranged from 9.5 to 10.5 in the two assemblies (Table 6), yielding one single variation per 100 bp in both cases.

	Radicchio CDS – 29,175 contigs			**Radicchio TEs – 122,745 contigs**		
	CHI-2418	**CHI-Witloof**	**Shared**	**CHI-2418**	**CHI-Witloof**	**Shared**
No. contigs	12,725	12,739	11,419	2,016	1,924	1,554
No. variants	123,843	121,086	119,729	10,662	10,651	10,246
No. variants/contigs	9.75	9.52	10.52	5.29	5.54	6.61
No. variants/100 bp	0.99 (1.14)	0.98 (1.14)	1.14 (2.05)	3.26 (3.64)	3.16 (3.50)	5.42 (8.88)
SNVs	115,678	113,049	107,255	9,532	9,605	9,006
MNVs	5,367	5,439	8,475	507	441	261
Insertions	2,044	2,036	2,166	556	552	714
Deletions	754	562	1,833	67	53	265

Table 6. Summary statistics of nucleotide variants restricted to gemonic regions of Radicchio annotated as CDS and Transposable Elements. Nucleotide variants were detected by using the transcriptomes CHI-2418 (wild type leaf chicory) and CHI-Witloof (cultivated Witloof type). For each transcriptome, the number of contigs displaying one or multiple variants, the number of variants and the number of variants per contigs are indicated. The number of variants per 100bp is also reported. Variants present in both transcriptomes are indicated as shared.

The vast majority of variants were Single Nucleotide Variants (SNVs), whereas Multi Nucleotide Variants (MNVs), Insertions, and Deletions were found to a considerably lower extent (Table 6). On average, the proportion of SNVs and MNVs was comparable in the CDS and TE contigs and equal to about 90% and 5%, respectively.

Among all contigs annotated as TEs, those characterized by the presence of one or multiple variants were 10,662 and 10,651 for the two transcriptomes (Table 6). The average number of variants per contig was equal to 5.3 and 5.5. Despite the relatively low abundance of polymorphic residues in these regions, the average number of variants per 100 bp was equal to 3.3 and 3.2. Single Nucleotide Polymorphisms (SNPs) were by far the most abundant type of variants in TEs as well as in CDS regions (Table 6). In particular, transversions and transitions were on average 37% (ranging from 35.6% and 37.8%) and 63% (ranging from 62.2% and 64.4%) of the point mutations, respectively. The total number of nonsynonymous SNPs calculated with the reference transcriptomes was equal to 13,559 (10.9%) and 11,197 (9.2%) for wild-type leaf chicory and cultivated Witloof accessions, respectively.

4. Discussion

Here, we report the uncovering of the first draft of the Radicchio genome. This highly relevant discovery was achieved by combining the recent advancement of next-generation sequencing technologies on the public side with the significant investment of financial resources in research and development on the private side.

Currently, conventional agronomic-based selection methods are supported by molecular marker-assisted breeding schemes. In recent years, we have demonstrated that the constitution of F1 hybrids is not only feasible in a small experimental scheme but also realizable and profitable on a large commercial scale (*e.g.*, registered CPVO varieties TT4070/F1, TT5010/F1, TT5070/F1, and TT4010/F1 in progress). F1 hybrids are varieties manifesting heterosis, or hybrid vigor, which refers to the phenomenon in which highly heterozygous progeny plants obtained by crossing genetically divergent inbred or pure lines exhibit greater biomass, faster speed of development, higher resistance to pests and better adaptation to environmental stresses than the two homozygous parents. Critical steps of an applicative breeding program are the production of parental inbreds. Two highly relevant factors in this context are the selection of self-compatible genotypes, to be used as pollen donors, and the identification of male-sterile genotypes, to be used as seed parents in large-scale crosses [21, 22].

It is worth mentioning that there are several reasons why the constitution of F1 hybrids is a strategic choice for a seed company. First, the crop yield of modern F1 hybrid varieties is usually much higher than that of traditional OP or synthetic varieties. Second, the uniformity of F1 populations and the way to legally protect their parental lines allow a seed company to adopt a plant breeder's rights, promoting genetic research and development programs that are very expensive and require many years. Finally, the need for breeding hybrid varieties also promotes the preservation of local varieties because the selection of appropriate inbred or pure lines as parents in pairwise cross-combinations requires the exploration and exploitation of germplasm resources. Our expectation is that F1 hybrid varieties will be bred and adopted with increasing frequency in Radicchio. Consequently, we invested in the sequencing and annotation of the first draft of the leaf chicory genome as it will have an extraordinary impact from both scientific and economic points of view. Indeed, the availability of the first genome

sequence for this plant species will provide a powerful tool to be exploited in the identification of markers associated with or genes responsible for relevant agronomic traits, influencing crop productivity and product quality. As an example, data and knowhow produced in this research project will be capitalized on in subsequent years to plan and develop basic studies and applied research on male-sterility and self-incompatibility in this species.

The availability of high-quality sequencing platforms (*i.e.*, Illumina) on the one hand, and specific and high-performing software for genome data assembly and gene set analysis on the other, made this project feasible. High-quality genomic DNA libraries were used for sequencing reactions performed with the Illumina platforms HiSeq and MySeq, originating a total of 197 million (mln) short reads and 29 mln longer sequences passing quality filters, respectively, which were then bioinformatically assembled to obtain the first genome draft. On the basis of this strategy, the genome draft of leaf chicory is composed of approximately 500,000 contigs, forming approximately 720 Mb. Based on the distribution of 25-mer frequencies, we estimated that the genome coverage is close to 25X. The same distribution also indicates that a significant part of the genome might be composed of highly repeated elements, as indicated by the number of k-mers that appears to be present with high frequency.

Nucleotide variant calling for the Radicchio genome showed comparable number of polymorphisms in the pairwise comparisons with the two publically available transcriptomes, originally developed from seedlings of two leaf chicory accessions (*i.e.*, wild and cultivated types). The total number of variants discovered in the CDS regions was shown to be approximately 10 times higher than the ones found in the TEs. This result might be a consequence of low expression, or silencing, of numerous transposable elements at the level of plant seedlings, as indicated by the finding that the mapping of the two transcriptomes to the reference genome failed to align sequences to about 98% of the contigs annotated as TEs. Noteworthy, the number of variations per 100 base pairs was significantly higher in the TEs than in the CSD sequences. This result might be explained by the accumulation of mutations in noncoding sequences, as most of the TEs are.

Overall, Single Nucleotide Variants (SNVs) were the most common variants compared with In/Del mutations. Since SNP mutations very often result in silent mutations, their high proportion in the CDS regions was an expected result. In/Del mutations that usually occur in silenced or functionally disrupted genes, along with noncoding regions, were found at a low rate in CDS regions.

TEs were found to occur, at least in one copy, in the 23.50% of the 522,301 contigs that constitute our chicory genome draft assembly. Retrotransposons proved to be the most abundant elements in the Radicchio genome. This finding is in agreement with data from previous studies [23-26]. It is worth mentioning that Copia-type elements were more abundant than Gypsy-type elements, forming the predominant subclass of LTR retrotransposons.

Although the amount of TEs of the totally assembled sequences was much lower than that reported for other species, the class ratio of the TE types corresponds to that found in previous studies [23-26]. Our estimate of TEs in leaf chicory is equal to 6.28% of the contigs length, which is much lower than amounts reported for soybean (59%), pigeonpea (52%), alfalfa (27%), trefoil

(34%), and chickpea (40%) [25, 27-30]. One of the reasons could be that our BLAST strategy chosen to find repeated elements in the genome was less efficient than specific software (*e.g.*, RepeatScout and RepeatMasker [31, 32]). Another reason could be the lack of TEs in the assembled portion of the Radicchio genome due to the low complexity of these repeated DNA regions.

The BLAST strategy with the nonredundant (NR) pentapetalae protein database produced the best output in terms of similarity with our contigs. This is undoubtedly due to the availability of large collections of sequences from species taxonomically related to leaf chicory, such as *Beta vulgaris*, *Helianthus annuus,* and *Lactuca sativa*, among others. Unfortunately, the depth of annotation of these recently sequenced genomes is frequently not comparable to that of the long-studied *Arabidopsis thaliana*. Although BLAST results obtained by querying the NR database proved to be highly informative in terms of the number of hits producing alignments with significant e-value, the annotation of the leaf chicory assembled contigs was more successful when the *A. thaliana* database was used alone. Therefore, a possible alternative for future enrichment of the current annotation state would imply the use of software (*e.g.*, Blast2GO) that could extract the annotation codes from multiple BLAST hits, provide the appropriate specificity cutoff, and assign the mapped GO terms to the original query.

Our choice to use the TAIR10 database to annotate our sequence contigs led to the annotation of a large number of assembled sequences and provided precious information concerning the putative process, or eventually, the metabolic pathways in which genes are putatively active.

The ability to annotate a certain number of sequences is not only exclusively dictated by the length and quality of the query sequences but also by their match with orthologous sequences that need to be annotated in depth.

This would be the case of annotations for metabolic pathways not actively studied or present in *A. thaliana* and for processes whose study is hampered by biological or physical circumstances. This might explain some discrepancies in annotations for male and female gametogenesis (Figure 6). From the graph, it is easy to understand the large discrepancy between the number of contigs presented for the term "Megagametogenesis" (GO:0009561), just 107, and the term "Pollen development" (GO:0009555), cited in the results as the most prevalent (more than six times that of megagametogenesis). We can suppose that this difference might not be due to a real difference in the number of genes involved in these two reproductive processes but rather to the lower number of genes known to be involved in female sporogenesis and gametogenesis.

Similarly, enzymes involved in the biosynthesis of germacren-type sesquiterpenoids, such as the germacrene-A synthase (EC:4.2.3.23), which are responsible for the biosynthesis of lactones associated with bitter taste in leaf chicory, are not known or properly characterized in *A. thaliana*.

Another fundamental finding of our study is the large number of SSR markers that were found in the assembled contigs. We can affirm that the leaf chicory genome shows an unexpected number and distribution of repeated sequences. Submitting our Radicchio draft to MISA software, we were able to reveal such a number of potential SSR markers. It is

therefore interesting that we were able to link a reasonably large number of microsatellites to each item here presented for both GO terms and KEGG maps. In the results, we presented only a small selection of important characteristics that could be utilized in marker-assisted selection and breeding programs in Radicchio. Together with SSRs, thousands of sequences that could be used in Single Nucleotide Polymorphism (SNP) analysis were associated to fundamental biosynthetic pathways or metabolism enzymes. This is a crucial starting point for modern breeding in leaf chicories.

It is noteworthy that further studies must be conducted to determine whether and how these potential markers could be exploited in molecular breeding programs. As a final step, gene prediction and annotation were also performed according to established computational biology protocols by taking advantage of the reference transcriptome data publically available for *Cichorium intybus* L. These sequences allowed us to learn the number, sequence, and role of the ~25.000 genes of the Radicchio's genome. This finding represents an important achievement for Italian agriculture genetics as a whole and opens new perspectives in both basic and applied research programs in Radicchio. It will have great impacts, potentials, and advantages in terms of breeding methods and tools useful for the constitution and protection of new varieties. Information obtained by the sequencing of the genome will be exploitable to detect and dissect the chromosomal regions where the genetic factors that control the expression of important agronomic and qualitative traits are located in Radicchio.

Modern marker-assisted breeding (MAB) technology based on traditional methods using molecular markers such as SSRs and SNPs, without relations to genetic modification (GM) techniques, will now be planned and adopted for breeding of vigorous and uniform F1 hybrids combining quality, uniformity, and productivity traits in the same genotypes.

In conclusion, our study will contribute to increase and reinforce the reliability of Italian seed firms and local activities of the Veneto region associated with the cultivation and commercialization of Radicchio plant varieties and food products; the seed market of this species will have the chance to become highly professional and more competitive at the national and international levels. To uncover the sequence of a given genome means to gain a robust scientific background and technological knowhow, which in short time can play a crucial role in addressing and solving issues related to the cultivation and protection of modern Radicchio varieties. In fact, we are confident that our efforts will extend the current knowledge of the genome organization and gene composition of leaf chicories, which is crucial in the development of new tools and diagnostic markers useful for our breeding strategies, and allow researchers for more focused studies on chromosome regions controlling relevant agronomic traits of Radicchio. In addition, conducting novel research programs for the preservation and valorization of the biodiversity, still present in the Radicchio germplasm of the Veneto region, is very important and accomplished through the genetic characterization of the most locally dominant and historically important landraces using sequenced genome information of Radicchio presented in this work.

Author details

Giulio Galla[1], Andrea Ghedina[1], Silvano C. Tiozzo[2] and Gianni Barcaccia[1*]

*Address all correspondence to: gianni.barcaccia@unipd.it

1 DAFNAE, Laboratory of Plant Genetics and Breeding, University of Padua, Legnaro, PD, Italy

2 T&T Produce, Sant'Anna di Chioggia, VE, Italy

References

[1] Lucchin M, Varotto S, Barcaccia G, Parrini P. Chicory and Endive. In: Springer Science, editor. *Handbook of Plant Breeding.* New York: 2008. p. 1-46. DOI: 10.1007/978-0-387-30443-4_1

[2] Barcaccia G, Pallottini L, Soattin M. Genomic DNA fingerprints as a tool for identifying cultivated types of radicchio (*Cichorium intybus* L.) from Veneto, Italy. *Plant Breed*. 2003;122:178-183. DOI: 10.1046/j.1439-0523.2003.00786.x

[3] Cadalen T, Morchen M, Blassiau C, Clabaut A, Scheer I, Hilbert JL *et al*. Development of SSR markers and construction of a consensus genetic map for chicory (*Cichorium intybus* L.). *Mol Breed*. 2010;25:699-722. DOI: 10.1007/s11032-009-9369-5

[4] Gonthier L, Blassiau C, Mörchen M, Cadalen T, Poiret M, Hendriks T, Quillet MC. High-density genetic maps for loci involved in nuclear male sterility (NMS1) and sporophytic self-incompatibility (S-locus) in chicory (*Cichorium intybus* L., Asteraceae). *Theor Appl Genet*. 2013;126(8):2103-2121. DOI: 10.1007/s00122-013-2122-9

[5] Muys C, Thienpont CN, Dauchot N, Maudoux O, Draye X, Van Cutsem P. Integration of AFLPs, SSRs and SNPs markers into a new genetic map of industrial chicory (*Cichorium intybus* L. var. *sativum*). *Plant Breed*. 2014;133(1):130-137. DOI: 10.1111/pbr.12113

[6] Ghedina A, Galla G, Cadalen T, Hilbert JL, Tiozzo CS, Barcaccia G. A method for genotyping elite breeding stocks of leaf chicory (*Cichorium intybus* L.). *BMC Research Notes*. Forthcoming.

[7] UC Davis. The Compositae Genome Project [Internet]. Available from: http://compgenomics.ucdavis.edu/cgp_wd_assemblies.php#13427 [Accessed: June 2015]

[8] Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull*. 1987;19:11-15.

[9] R. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821-829. DOI: 10.1101/gr.074492.107

[10] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-477. DOI: 10.1089/cmb.2012.0021

[11] Cornell University Library - Binghang Liu, Yujian Shi, Jianying Yuan, Xuesong Hu, Hao Zhang, Nan Li, Zhenyu Li, Yanxiang Chen, Desheng Mu, Wei Fan. Estimation of genomic characteristics by analyzing kmer [Internet]. 2013. Available from: http://arxiv.org/abs/1308.2012

[12] NCBI [Internet]. Available from: http://www.ncbi.nlm.nih.gov/

[13] UniProt [Internet]. Available from: http://www.uniprot.org/

[14] Gene Ontology Consortium [Internet]. [Updated: http://geneontology.org/].

[15] KEGG: Kyoto Encyclopedia of Genes and Genomes [Internet]. Available from: http://www.genome.jp/kegg/

[16] Blast2GO [Internet]. Available from: https://www.blast2go.com/

[17] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005;21(18):3674-3676. DOI: 10.1093/bioinformatics/bti610

[18] Galla G, Barcaccia G, Ramina A, Collani S, Alagna F, Baldoni L, Cultrera N. GM, Martinelli F, Sebastiani L, Tonutti P. Computational annotation of genes differentially expressed along olive fruit development. *BMC Plant Biol*. 2009;9(128). DOI: 10.1186/1471-2229-9-128

[19] MISA - MIcroSAtellite identification tool [Internet]. [Updated: 5/14/02]. Available from: http://pgrc.ipk-gatersleben.de/misa/

[20] Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH). PGSB Plant Genome and Systems Biology [Internet]. Available from: http://pgsb.helmholtz-muenchen.de/plant/recat/

[21] Barcaccia G, Tiozzo CS. New male sterile Cichorium spp. mutant, parts or derivatives, where male sterility is due to a recessive nuclear mutation linked to a polymorphic molecular marker, useful for producing F1 hybrids of Cichorium spp. EU PATENT No. WO2012163389-A1. 2012;DOI: 1013140/2.1.4749.5044

[22] Barcaccia G, Tiozzo CS. New male sterile mutant of leaf chicory, including Radicchio, used to produce chicory plants and seeds with traits such as male sterility exhibiting cytological phenotype with shapeless, small and shrunken microspores in dehiscent anthers. USA PATENT No. US20140157448-A1. 2014;DOI: 10.13140/2.1.3176.6400

[23] Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, Cahoon EB, Gedil M, Stanke M, Haas BJ, Wortman JR, Fraser-Liggett CM, Ravel J, Rabinowicz PD. Draft genome sequence of the oilseed species *Ricinus communis*. *Natur Biotechnol*. 2010;28:951-956. DOI: doi:10.1038/nbt.1674

[24] Rahman AYA, Usharraj AO, Misra BB, Thottathil GP, Jayasekaran K, Feng Y, Hou S, Ong SY, Ng FL, Lee LS, Tan HS, Sakaff MKLM, Teh BS, Khoo BF, Badai SS, Aziz NA, Yuryev A, Knudsen B, Dionne-Laporte A, Mchunu NP, Yu Q, Langston BJ, Freitas TAK, Young AG, Chen R, Wang L, Najimudin N, Saito JA, Alam M. Draft genome sequence of the rubber tree *Hevea brasiliensis*. *BMC Genomics*. 2013;14(75). DOI: 10.1186/1471-2164-14-75

[25] Jain M, Misra G, Patel RG, Priya P, Jhanwar S, Khan AW, Shah N, Singh VK, Garg R, Jeena G, Yadav M, Kant C, Sharma P, Yadav G, Bhatia S, Tyagi AK, Chattopadhyay D. A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.). *Plant J*. 2013;74:715-729. DOI: 10.1111/tpj.12173

[26] He N, Zhang C, Qi X, Zhao S, Tao Y, Yang G, Lee T-H, Wang X, Cai Q, Li D, Lu M, Liao S, Luo G,He R, Tan X, Xu Y, Li T, Zhao A, Jia L, Fu Q, Zeng Q, Gao C, Ma B, Liang J, Wang X, Shang J, Song P, Wu H, Fan L, Wang Q, Shuai Q, Zhu J, Wei C, Zhu-Salzman K, Jin D, Wang J, Liu T, Yu M, Tang C, Wang Z, Dai F, Chen J, Liu J, Zhao S, Lin T, Zhang S, Wang J, Wang J, Yang H, Yang G, Wang J, Paterson A.H, Xia Q, Ji Q, Xiangc Z. Draft genome sequence of the mulberry tree *Morus notabilis*. *Natur Commun*. 2013;4. DOI: 10.1038/ncomms3445

[27] Sato S, Nakamura Y, Kaneko T *et al*. Genome structure of the legume, *Lotus japonicus*. *DNA Res*. 2008;15(4):227-239. DOI: 10.1093/dnares/dsn008

[28] Schmutz J, Cannon SB, Schlueter J *et al*. Genome sequence of the palaeopolyploid soybean. *Nature*. 2010;463:178-183. DOI: 10.1038/nature08670

[29] Varshney RK, Chen W, Li Y *et al*. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Natur Biotechnol*. 2012;30:83-89. DOI: 10.1038/nbt.2022

[30] Young ND, Debelle F, Oldroyd GE *et al*. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature*. 2011;480:520-524. DOI: 10.1038/nature10625

[31] Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics*. 2005;21:351-358. DOI: 10.1093/bioinformatics/bti1018

[32] Smit AFA, Hubley R, Green P. Repeat Masker Open-4.0 [Internet]. 2013-2015. Available from: http://www.repeatmasker.org