

Université catholique de Louvain



Earth and Life Institute – Agronomy, crop physiology and plant breeding



Genetic determinism of inulin metabolism in industrial chicory (*Cichorium intybus* L.): an association mapping approach.

Pierre Raulier
November 2015

Promoters:	Prof. Draye Xavier (UCL/ELI/ELIA) Prof. Pierre Bertin (UCL/ELI/ELIA)
President:	Prof. Bruno Hendy de Frahan
Members of the jury:	Prof. Marc Boutry (UCL/ISV) PhD Hilde Muylle (ILVO/ Plant Sciences Unit) Ir. Christine Notté Prof. Claire Perilleux (ULG/ PhytoSYSTEMS) Prof. Pierre Van Cutsem (UNamur/URBV)

Remerciements

Je commencerai bien sûr en adressant mes remerciements à mes chers promoteurs : Pierre Bertin et Xavier Draye.

Pierre, tu as guidé mes premiers pas sur les chemins de l'expérimentation scientifique, l'écophysiologie et l'analyse de données, le tout parsemé d'une touche de génétique! Tu m'as ensuite donné la confiance pour me plonger corps et âme dans une carrière scientifique avec un premier projet sur le chanvre. Merci!

Suite à un tour de tour de passe-passe dont seul le destin à le secret, je me suis retrouvé embarqué dans un projet traitant non plus du chanvre mais de la chicorée, le tout avec plus de génétique et non plus un, mais deux promoteurs! Xavier, avec toi j'ai eu le bonheur d'explorer les tréfonds de la génétique quantitative et de l'analyse de donnée, miam et Merci!

Pierre et Xavier, vous m'avez offert la plus belle atmosphère de travail dont on puisse rêver au sortir des études : confiance, bienveillance et inventivité. Pour tout ça, Merci !

La liste des mentors se poursuit naturellement avec Anne-Michelle qui m'a accueilli comme mémorant, collègue et enfin ami ! Chef, bosser avec toi, c'était super!

Chicoline est à l'origine du projet INULONG dans lequel s'inscrit cette thèse. Chicoline a ensuite rendu ce projet possible grâce à un magnifique travail de fourmis! Fourmis multidisciplinaires d'ailleurs ! Des fourmis moléculaires, capables de géotyper 1497 individus diploïdes soit un total de 2 996 allèles, des les fourmis en bottes qui, sans relâche et avec le plaisir du travail bien fait, ont semé, dépoté, planté, bichonné, récolté, rebichonné et, finalement, un maltraité, un total de 14 000 racines. D'autres fourmis en blouse blanche ont explosé le record MONDIAL de profilage de fructanes chez la chicorée! Tout ce travail a bien sûr été réalisé en plus des travaux de routine et d'agrandissement de la fourmilière Chicoline !

Bref, un grand merci à l'équipe de terrain : Veronique, Gery, Gregory et Philippe. Je garderai longtemps en mémoire votre accueil et l'ambiance détendue mais efficace qui règne sur le terrain!

Un autre tout grand Merci a Vinciane et Virginie pour ces nombreuses journées semaines, mois à analyser ADN et inuline !

Christine, Olivier et Vincent, je resterai impressionné par la passion et le savoir qui vous lie à vos métiers, ainsi qu'en vos qualités de transmission ! Un grand merci pour votre accueil !

Nicolas Dauchot, collègue direct sur la ligne de front d'INULONG, j'ai apprécié la dynamique de travail que nous avons créé: l'un initie une bride de séquençage, l'autre entreprend le séquençage complet de la bête. L'un produit des données, l'autre les analyse! Ces interactions actives impliquant nos spécialités respectives

auront fait partie des travaux les plus agréables de cette thèse. Au plaisir donc!

En plus de ces nombreux collaborateurs, j'ai eu la chance d'avoir un jury multidisciplinaire et attentif. Merci donc à Christine Notté, Hilde Muylle, Claire Périlleux, Marc Boutry, Pierre Van Cutsem pour vos discussions et remarques avisées.

Passons maintenant à la famille ECAV ! Avec Isabelle, qui, telle une bergère, guide mémorants, doctorant et académiques au travers des méandres de la machinerie administrative UCL ! Accueillante et toujours à l'affût d'une petite causette, je garderai longtemps en tête son joyeux « PETARD » qu'elle lâche pour exprimer une petite contrariété!

Merci à Willy, sans qui, bien, on ne sait pas ce que l'on serait devenu aujourd'hui !

Un tout grand merci à Pierre et Thomas, pour leur bonne humeur, leur efficacité et leur inventivité lors de nos petites expéditions hainuyères!

Nicolas et Nicolas, deux autres acolytes qui se sont chacun aventurés hors des sentiers battus d'INULONG pour explorer les formes racinaires ! Merci pour votre ingéniosité et votre enthousiasme !

Collègues doctorants et postdocs, quel plaisir d'avoir pu, ces 6 dernières années, partager avec vous ces 1678 pauses-café. J'ai aussi usé 1,72 chaise de bureau en votre présence tiens ! Heureux, donc d'avoir commencé ma thèse avec les

charmantes Aurélie, Anne-Mich, Beata, Inès et Laure-Anne, ainsi qu'avec ces bons vieux Guillaume, René et John! Toujours content aussi, quand l'un d'entre vous vient retrouver ses marques dans le B231! L'équipe de départ a été très agréablement agrémentée par Hahn, Michelle, Nathalie, Alek, Igor, Léopold, Maxime et Yves. J'aurai eu l'honneur de terminer avec un Crew du tonnerre composé de quelques coquins toujours à l'affut d'une blagounette : Adrien, Ben, Gauthier, Marc et René!

Un merci particulier à Guillaume et Xavier pour m'avoir embarqué dans ce qui allait devenir ma grande passion: la grimpe ;)

La vie à l'université ne serait pas complète sans ces nombreux petits casse-croustes avec Julie, Viviane, Nicof, Raphaël, les collègues et bien d'autres! Je salue également le slack crew (Marie-Liesse, JB, Seb, Vinc, ...) pour ces petites escapades ensoleillées !

Big-up à tous les colocos qui m'ont accompagné ces 6 années de thèse. KJNistes, lapins ou brochetons, vivre avec vous était un plaisir quotidien ! Un merci particulier à Adèle, Aurore, Antoine, Coco, Stanley, Anne-Cath et Rob qui m'ont assuré un soutien logistique et moral lors de ces derniers mois de rédaction!

Un gros merci à Anne-Cath, Maria et Ben pour votre sérieux coup de main dans les moments de rush !

Coucou à tous les copains que je n'ai pas pu saluer ici !

Enfin, les dernières pensées iront à ma famille,
à Maman, Papa et Céline, pour le bonheur que nous avons de nous réunir,
à Isabelle, Xavier, Gébou et Nical pour m'avoir offert une place particulière dans
la vie de Camille et Zonzon.

Table of contents

1	Introduction and objectives	1
2	Literature review	5
2.1	Chicory	7
2.1.1	Botany	7
2.1.2	A brief overview of history of industrial chicory	10
2.2	Fructans	13
2.2.1	Nomenclature	13
2.2.2	Distribution across species	16
2.3	Metabolism	17
2.3.1	Fructan metabolism enzymes	17
2.3.2	Origin of the differences in fructan quality and concentration.	23
2.3.3	The dynamic of fructans accumulation in chicory is divided in three phases.....	24
2.4	Fructan profile in chicory: an ideotype.....	33
2.5	Genetic determinism of fructans.....	36
2.5.1	QTL involved in fructan quality has been identified in different species. 36	
2.5.2	Chicory presents exploitable genetic variability	38

2.6	Reference	41
3	An introduction to population and quantitative Genetics.	51
3.1	Population genetics.....	54
3.1.1	Genetic diversity	54
3.1.2	Population structure	58
3.1.3	Linkage disequilibrium	72
3.2	Linking genetic factors to phenotypes	83
3.2.1	Phenotype determinism	83
3.2.2	QTL mapping and GWAS: an overview.....	85
3.2.3	GWAS	90
3.3	References	98
4	Thesis outline	101
4.1	General objective.....	103
4.2	Specific objectives.....	103
4.3	Strategy.....	104
4.4	Thesis organization.....	111
5	Exploration of genetic diversity within Cichorium endivia and Cichorium intybus with focus on the gene pool of industrial chicory.....	117
5.1	Abstract	119
5.2	Introduction	121

5.3	Materials and methods.....	128
5.3.1	Plant material	128
5.3.2	DNA extraction and SSR analyses	136
5.3.3	Population analyses	136
5.4	Results	138
5.4.1	SSR polymorphism (first and second panel)	138
5.4.2	Genetic diversity among cultivated <i>Cichorium</i> (first panel)	142
5.4.3	Genetic diversity among modern industrial cultivars (second panel) 151	
5.5	Discussion	157
5.5.1	Species delimitation	157
5.5.2	Relationships among the three <i>C. intybus</i> cultivar groups	158
5.5.3	‘Magdeburg’, Witloof and modern industrial cultivars	159
5.5.4	Leaf cultivars.....	161
5.5.5	Wild individuals	161
5.5.6	Modern industrial cultivars	162
5.6	Conclusion and perspectives	164
5.7	References	166
5.8	Supplementary materials	170
6	Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest inulin depolymerization	175

6.1	Abstract	177
6.2	Introduction	178
6.3	Material and methods	182
6.3.1	Sample sets.....	182
6.3.2	Plant Material and DNA extraction.....	183
6.3.3	Carbohydrate characterization.....	183
6.3.4	Primer design and PCR amplification	184
6.3.5	EcoTILLING.....	187
6.3.6	DNA sequencing	188
6.3.7	Genotyping.....	188
6.3.8	Data analysis	189
6.4	Results	190
6.4.1	Polymorphism and association study	190
6.4.2	Confirmation of the association	199
6.5	Discussion	203
6.6	Conclusions	208
6.7	Reference.....	209
6.8	Supplementary materials	215
7	Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in Cichorium intybus than copy number variation of its close paralog 1-FEH IIa.	225

7.2	Abstract	227
7.3	Introduction	228
7.4	Materials and Methods	232
7.4.1	Plant Material	232
7.4.2	Root sampling and carbohydrates phenotyping.....	232
7.4.3	Leaf sampling and DNA extraction.....	233
7.4.4	PCR and qPCR primer design	234
7.4.5	qPCR reaction mixture and cycling conditions	236
7.4.6	qPCR efficiency	236
7.4.7	qPCR	238
7.4.8	Cloning and sequencing of 1-FEH I and 1-FEH IIb genomic regions	239
7.4.9	Data analysis	241
7.4.10	Statistical tests	241
7.5	Results	242
7.5.1	Normalization genes for qPCR in chicory.....	242
7.5.2	Specificity of qPCR primer pairs	243
7.5.3	qPCR efficiencies	243
7.5.4	Association of GH32 NRQ data with post-harvest inulin depolymerization.....	244
7.5.5	Identification of a loss-of-function allele of 1-FEH IIb	251

7.5.6	Impact of the loss of mini-exon 2 in FEH IIb	255
7.6	Discussion	257
7.7	Conclusions	267
7.8	Reference	269
7.9	Supplementary materials	275
8	GWAS analysis of carbohydrate and inulin quality in chicory	295
8.1	Abstract	297
8.2	Introduction	298
8.3	Materials and methods.....	302
8.3.1	Plant material	302
8.3.2	Phenotyping.....	304
8.3.3	Phenotypic analysis.....	306
8.3.4	Genome sequencing	307
8.3.5	Genotyping-by-Sequencing.....	308
8.3.6	Population analysis.....	309
8.3.7	Linkage disequilibrium	311
8.3.8	GWAS	311
8.4	Results	313
8.4.1	Reference genome	313
8.4.2	SNPs discovery and distribution	313

8.4.3	Population structure and linkage disequilibrium	316
8.4.4	Phenotyping.....	318
8.4.5	GWAS	324
8.5	Discussion	331
8.6	References	340
8.7	Supplementary materials	348
9	Discussion and conclusion	357
9.1	Work package 1: Characterization of the genetic diversity of industrial chicory	359
9.2	Work package 2: Creation of a diversity panel suitable for candidate-gene and genome-wide studies	361
9.3	Work package 3: Phenotyping	362
9.4	Work package 5: Genotyping	364
9.5	Work packages 4 & 6: identification and validation of candidate genes	365
9.6	Work package 7: Genome wide investigation of genetic factors associated to variation on carbohydrate synthesis and cold-induced inulin degradation.....	369
9.7	General conclusion	372
9.8	Reference	373

Acronyms

AFLP: Amplified fragment length polymorphism

AMOVA: Analysis of molecular variance

ANOVA: Analysis of variance

bp : base pair

DP: polymerization degree

Dps: distance proportion of shared alleles

Fst: Fixation index

GBS: genotyping-by sequencing

GWAS: genome-wide association mapping

ITS: Internal transcribed spacer

LD: Linkage disequilibrium

mDP: mean polymerization degree

PIC: Polymorphism index content

PCA: Principal component analyses

SSR: Simple sequence repeats

SNP : Single nucleotide polymorphism

UPGMA: Unweighted pair group method with arithmetic mean

UTR: UnTranslated Region

1-SST: sucrose:sucrose 1-fructosyl-transferase

1-FFT: fructan:fructan 1-fructosyltransferase

6-SFT: sucrose:sucrose-1 fructosyltransferase

6G-FFT: fructan-fructan 6-G-fructosyltransferase

FEH: fructan exohydrolases

1 Introduction and objectives

Inulin is a polymer of fructose used as prebiotic and dietary fiber. It is mainly extracted from the root chicory. The quality of the chicory root extract is defined by the polymerization degree of inulin molecules (number of fructose units) and the concentration of free fructose. The processing industry currently values extracts with a high inulin polymerization degree (Van Arkel et al., 2013) and with low concentration of free fructose.

The polymerization degree and the free fructose content result from the equilibrium between inulin synthesis and degradation. This equilibrium is affected by environmental conditions, especially by low temperatures prevailing during the harvesting period -from September to December- in Belgium. Low temperatures initiate inulin degradation, resulting in a decrease of the inulin yield and polymerization degree and an increase of the undesirable free fructose. Reducing the extent of the cold induced degradation is one of the current breeding targets, but relies on extensive, costly and destructive measurements

Breeders aim to create cultivars matching industrial requirements. Since the revision of the European sugar regime in 2006, occurrence of free fructose in the root extracts has become undesirable and the breeders have increased their focus on inulin quality. Given the long-term vision needed in crop improvement programs, it is the interest of breeders to develop strategies in order to match the rapidly evolving industrial quality standards.

In this context, the general objective of this thesis is to unravel the genetic determinism of the synthesis of carbohydrate (including inulin) and the cold-induced inulin degradation kinetics occurring in field. This will enable the breeder to develop marker-based selection strategies in order to reach both current and future breeding objectives.

2 Literature review

2.1.1 Botany

Chicory (*Cichorium intybus* L.) is a plant of the *Asteraceae* family naturally present in Europe, Northern Africa, Middle East Asia, and Iran (Agro Atlas 2015). This biannual species requires vernalisation to initiate bolting. Wild chicory is allogamous and sensitive to inbreeding depression. Self hybridization is prevented by partial self-incompatibility, pollen competition and protandry (Eenink 1981). Chicory is a diploid species ($2n = 18$) and the the genome size is estimated to 1300 CM and between 600 to 1300 cM (Koopman 2002, Cadalen et al. 2010; Gonthier et al. 2013; Muys et al. 2014; Simone et al. 1997; Van Stallen et al. 2003)

Cultivated chicory is an interesting model of morphological diversity. The work of generations of farmers led to the creation of three contrasted groups of cultivars (Kiers et al. 2000). The group of leaf chicory is composed of green or red leafy vegetables that can be consumed fresh or cooked. This cultivar group presents an important diversity in shape and color and includes three cultivars groups: the Raddichio group, the Sugarloaf group and the Catalogne group (Figure 1). The group of Witloof chicory consists in white vegetables made up of tightly packed leaves after forcing under artificial condition. The group of root chicory, also known as industrial chicory, is characterized

by a large tap root. It was originally grown for coffee substitute production and is also now grown for inulin production. Inulin is the main chicory product in Belgium. Chapter 2 provides more details concerning these different groups.

Apart from disease resistances, the genetic improvement of the three different cultivar groups is divergent. Leaf chicory breeders are currently improving gustative, shape and color characteristics (Ryder, 1999). Witloof breeders are looking towards the uniformity of the shape, the quality and the yield of forced leaves. Finally, industrial chicory breeders focus on root shape, early vigor, bolting resistance, inulin quality and yield (Doré & Varoquaux, 2006). Originally, cultivars of the different cultivated forms were all open-pollinated cultivars. For more than 10 years, the breakdown of auto-incompatibility systems has enabled the creation of inbred lines and F1 hybrids in all cultivar groups.

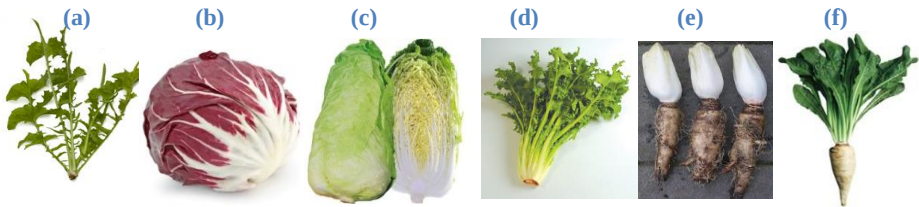


Figure 1: Chicories from different cultivar groups (a) wild individual (b) leaf chicory, Radicchio type (c) leaf chicory, Sugarloaf type, (d) leaf chicory, Catalogne type, (e) witloof, (f) industrial chicory.

The chicory taproot is thin and unregular in wild chicories and leaf cultivars and rather large and conical in Witloof and industrial chicories. In all case, the secondary growth is result of the action of one or possibly two cambium layer (Wieser, 1965) (Figure 2). Since the cambium is the structure responsible for the secondary grow of the roots, we may expect that the potential of secondary grow in chicory is limited compared to plants such as sugar beet (*Beta vulgaris* L.) that present multiple cambium layers. Sucrose produced by the arial part is transported by the phloem, in the parenchima cells of the roots. There held most of the fructan synthesis. Parenchima includes many laticifère producing latex (Esau, 1965).

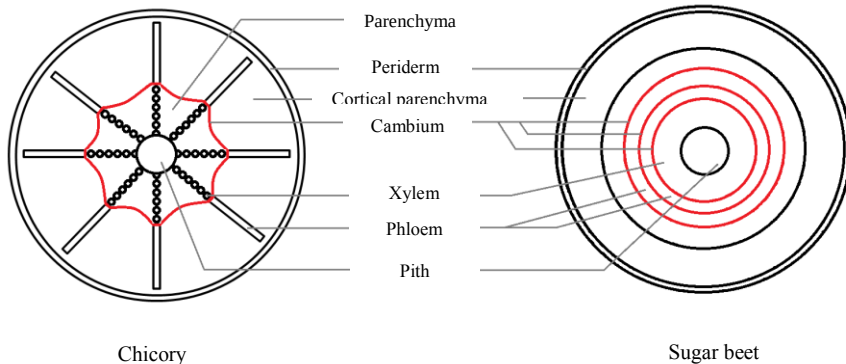


Figure 2: transversal cut of chicory (left) and sugar beet (right).

2.1.2 A brief overview of history of industrial chicory

During the XVIIIth century, the Napoleon blockade led to a drastic increase of coffee prices and the roasted root of chicory became a popular coffee substitute. During this period, the chicory cultivation developed in Belgium, the north of France, Germany and Austria. Coffee substitute remained the main use of industrial chicory till the 1990's. Originally, root chicories were open-pollinated cultivars characterized by a long taproot, prone to manual harvest (Figure 3). The mechanical harvest appeared during the seventies. At this time, the long taproots were often broken by mechanical harvest, so that a substantial part of the root production remained into the soil. This led the breeders to develop cultivars better adapted to mechanical harvest, characterized by conical taproots with an increased root biomass in the topsoil. Tilda, the first cultivar adapted to mechanical harvest was released in 1981.

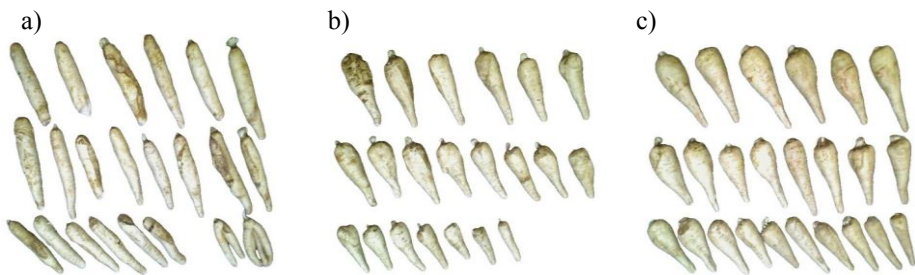


Figure 3: Roots of 3 contrasted chicory cultivars. a) Novipa, released in 1956, is a typical ancient cultivar adapted to manual harvest. It is characterized by a long cylindrical root. b) Tilda, released in 1981, is the

first cultivar adapted to mechanical harvest. It presents a short, massive conic root. c) Melci, released in 2001 is much bigger and homogenous than Tilda.

Inulin industry started in Belgium and The Netherlands in the second half of the eighties, with the reconversion of sugar refinery into chicory refinery in Warcoing (Cosucra group Warcoing, 1986), in Oreye (Beneo-Orafti SA, 1991) and Roosendaal (Sensus a sub group of the Royal Cossun Company, 1992). Those factories are still active today. During this period, industrial chicory products were inulin and fructose syrup. Inulin is a fructose polymer used as probiotic, dietary fiber, low calories sweetener or fat substitute. It is characterized by a sweet taste when constituted of a low number of fructosyl units and has no taste and a greasy aspect when the number of fructosyl is above 10 (Franck, 2008). Fructose syrup can be obtained after hydrolysis of inulin. In Belgium, the inulin production and chicory surface increased after the opening of the second factory and reached a maximum of 16 000 ha in 2004 (FAO, 2015) (Figure 4).

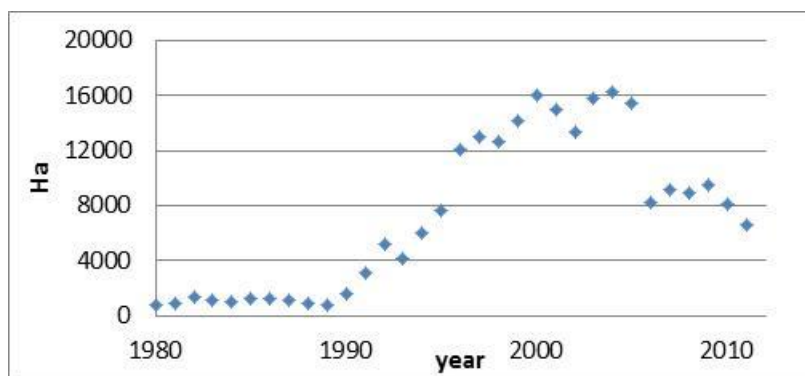


Figure 4: Evolution of chicory culture surfaces in Belgium.

The development of cultivars dedicated to inulin production started with the inulin industry. Different companies were involved in chicory breeding, two of which are still active: Chicoline, a sub-group of Cosucra group Warcoing, and Florimond-Deprez. Breeders initiated their selection by improving the bolting resistance, disease resistance, dry yield and root shape of “coffee” cultivars. The breeding strategies are designed to take into account the outcrossing nature of chicory, the limited self-hybridation, the important inbreeding depression of the individuals from self hybridation and the destructive nature of the carbohydrate characterization. Besides, the cloning of chicory is relatively easy but results in plants with disturbed root system. The cloned individuals are useful for the characterization of traits associated to aerial part and for the production of new crosses. Therefore, industrial chicory is largely bred by crosses, multi-testing, progeny testing and recurrent selection. Due to the outcrossing nature of chicory, most of the cultivars released on the market are synthetic cultivars stemming from the cross of a set of selected individuals. The development of male sterile lines by protoplast fusion of chicory and Sunflower (*Helianthus tuberosus* L.) also allowed the development of hybrid cultivars since 1988 (Doré & Varoquaux, 2006).

The chicory market was drastically affected by the 2006 revision of the European sugar regime as inulin syrup was withdrawn from the market and fructose became an inulin byproduct. The European Union supported the chicory industry to adapt the refineries to inulin extraction only. This led the breeders to orient their selection toward low fructose

cultivars. This change was further motivated by the fact that fructose hampers industrial inulin extraction. Another consequence of this reform is the reduction of the cultivated surfaces, as shown on the Figure 4.

2.2 Fructans

2.2.1 Nomenclature

Fructans are oligo- or polysaccharides that contain at least two adjacent fructosyl units. Different types of fructans can be distinguished, according to (i) the type of fructosyl-fructose linkage, (ii) the occurrence and position of the glucose unit (G) and (iii) the number of fructosyl units (F).

Inulin-type fructans consist in fructose polymers mostly or exclusively linked with β (2 $\rightarrow$ 1) fructosyl-fructose linkage whereas the **levan**-type fructans are mostly or exclusively linked with β (2 $\rightarrow$ 6) fructosyl-fructose linkage. Both inulin and levan fructose chains are terminated by a terminal glucose ($GF_n, n \geq 2$). These fructans are essentially linear molecules but a low degree of branching may occur when the two kinds of linkage are present in the same molecule. When the fructose chain lacks the terminal glucose, the terms **inulo-*n*-ose** [β (2 $\rightarrow$ 1)] and **levan-*n*-ose** [β (2 $\rightarrow$ 6)] are used ($F_n, n \geq 2$). The 1-kestose and the 6-kestose are the smallest representatives of inulin and levan type

respectively and both include one glucose and two fructosyl units (Figure 5). The corresponding inulo-*n*-ose and levan-*n*-ose molecules are the inulobiose and levanbiose, which both consist in two fructose molecules. The **bifurcose** is the smallest mixed fructan. It consists of one glucose and tree fructosyl units involving β (2→1) and β (2→6) link.

Some fructans consist in fructosyl units flanking both sides of a glucose unit ($F_nGF_m, n \text{ \& } m \geq 2$). These **fructan neoserries** comprise **inulin neoserries** and **levan neoserries**, according to the type of fructosyl-fructose linkage. The smallest representative of the fructans neoserie is the **neokestose** that consists in one glucose unit surrounded by two fructosyl units.

The **degree of polymerization** (DP) characterizes the number of fructosyl units of a fructan molecule. As natural extracts are constituted of fructan molecules with different DP, they are characterized by their mean DP (mDP), which estimates the number of fructose units of the “average” fructan molecule of the extract.

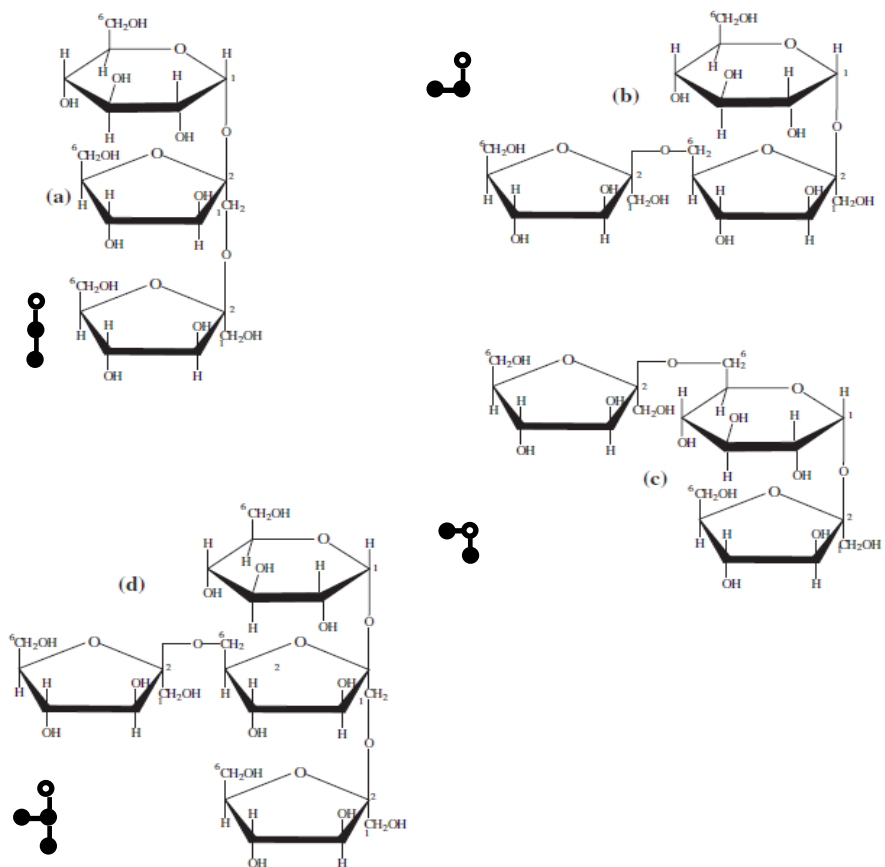


Figure 5: Formulae of some fructan oligosaccharides. (a) 1-Kestose; (b) 6-Kestose; (c) neokestose; (d) bifurcose (adapted from Van Laere and Van den Ende 2002).

2.2.2 Distribution across species

Fructans are produced by approximatively 15 % of plant species from many genera (Hendry 1993). Fructans, like starch, are primarily a storage substance. They also contribute to membrane protection during physiological stresses such as cold or water deficit, through the regulation of the cell osmotic potential (Pilon-Smits et al. 1995; Pollock 1986). Accordingly, fructan plants are generally native from temperate regions with seasonal drought and frost periods, and are generally absent in tropical regions and in wetlands (Hendry 1993).

The type of fructan, its mDP and its distribution in the plant and its concentration vary among species. **Dicot** species mainly produce inulin and inulose while **monocots** produce a wider range of fructan types. Among dicot species producing fructans, chicory synthesizes and accumulates fructans (mDP~12) in the roots, Jerusalem artichoke (*Helianthus tuberosus* L.) synthesizes long fructans (mDP > 100) in the aerial part and accumulates them in the tuber while globe artichoke (*Cynara scolymus* L.) synthesizes and accumulates long fructan (mDP~50) in the flowers. Inulin represents up to 95 % of the dry weight of the chicory root and Jerusalem artichoke tuber, against 40 % of the dry weight of globe artichoke flowers (Hellwege et al. 1998; Van Laere and Van den Ende 2002; López-Molina et al. 2005; Pollock 1986).

In monocots, Poaceae generally accumulate several types of fructan (inulin, inulin neoserries, levan, levan neoserries, ...) while Liliaceae

such as onion and asparagus store inulin and levan neoseries (Livingston, Chatterton, and Harrison 1993; Shiomi 1989).

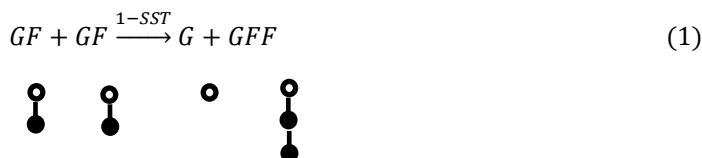
2.3 Metabolism

2.3.1 Fructan metabolism enzymes

The metabolism of fructans has been studied in the past decades in a number of plant species, with emphasis on grasses, cereals and chicory. These studies deal with many aspects of fructans metabolism, including the identification and characterization of substrates, enzymes and genes. The metabolism of fructans is mediated by enzymes of the family 32 of glycosyl hydrolases (GH32). These enzymes are all involved in the transfer reactions of fructosyl units and are primarily localized in the vacuole. They differ by their affinity for, and the type of donor and acceptor molecules. Their activity depends on a number of factors, including species, phenology, tissue, time and environmental conditions.

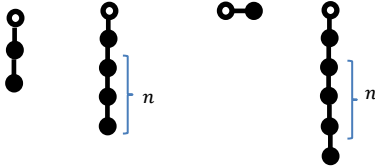
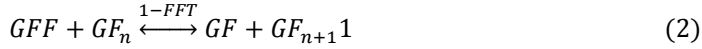
Fructans biosynthesis is mediated by four enzymes that primarily catalyze two types of reactions. The **initiation** reaction consists in the transfer of a fructosyl unit from a sucrose (GF) to the fructosyl unit of another sucrose in order to create an initial trisaccharide fructan (GFF), while the **elongation** reaction consists in the addition of fructosyl units to an existing fructan chain.

The **sucrose:sucrose 1-fructosyl-transferase (1-SST)**, found in dicots and some monocots, transfers a fructose moiety from a sucrose molecule to the C-1 of the fructose of another sucrose molecule (initiation step), creating a β (2 $\rightarrow$ 1) glycosidic bond (eq 1). This essentially irreversible reaction yields a trisaccharide 1-kestose and a free glucose (Edelman and Jefford 1968) .

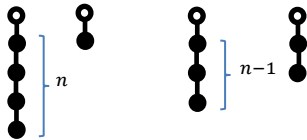
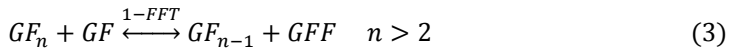


The **fructan:fructan 1-fructosyltransferase (1-FFT)**, found in dicots and some monocots, transfers a fructosyl unit from a donor molecule to the C-1 of an acceptor molecule creating a β (2 $\rightarrow$ 1) glycosidic bond. The donor or acceptor molecule can both be a sucrose or a fructan. Most fructan molecules can act as donor or acceptor in this reaction, depending on their relative concentration. However, the nature of the donor and acceptor molecules of the 1-FFT transfer reaction will affect the resulting fructan type and the mDP of the fructan pool. This confers 1-FFT a crucial role in the fructan pool constitution. In the presence of high concentrations of 1-kestose, the 1-FFT reaction uses 1-kestose as donor, removes it from the inulin pool generate a sucrose and increases the DP of the acceptor, leading to an increase of the mDP (eq. 2). In this context, 1-FFT can act synergistically with 1-SST as 1-SST regenerates the initial 1-kestose from the sucrose released by the 1-FFT reaction

(Edelman and Jefford 1968).

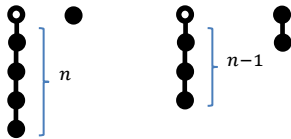
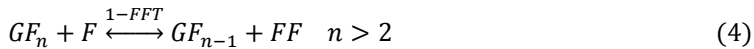


On the opposite, in limiting concentration of 1-kestose, 1-FFT uses a longer fructan chain as donor (eq. 3). This reaction will not affect the mDP of the inulin pool, except if the acceptor is a sucrose molecule. In the latter case, the transfer reaction generates two fructan molecules from one sucrose molecule and one inulin molecule. This **back-transfer** reaction introduces an equal amount of glucose and fructose in the inulin pool, increases the inulin pool and reduces the mDP (Wim Van den Ende and Van Laere 1996a, 1996b).



Finally, the 1-FFT can transfer a fructosyl unit from an inulin molecule to a free fructose (when in excess) or fructose polymer (inulose), leading to the synthesis of a new inulose molecule or to the elongation of an existing inulose and to the shortening of the fructan donor. The

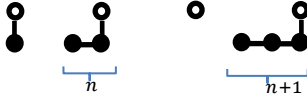
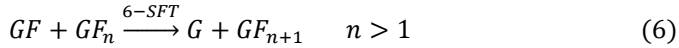
consequence is a negative or null effect on the mDP of the fructan (inulin and inulose) pool (eq 4) (Wim Van den Ende and Van Laere 1996a, 1996b).



The **sucrose:fructan 6-fructosyltransferase (6-SFT)**, found in monocots only, transfers a fructose moiety from a sucrose molecule to the C-6 of a fructose of others molecule, producing a β (2→6) glycosidic bond. Three types of transfer may be achieved by 6-SFT depending on the nature of the acceptor. When the acceptor is the fructose of a sucrose molecule, the transfer reaction produces a free glucose and **initiates** a new fructan molecule (6-kestose, eq 5).



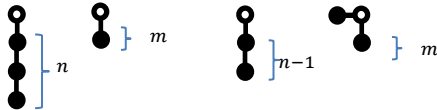
When the acceptor is the final fructose of an inulose or levanose, the transfer reaction produces a glucose and **elongates** the fructan molecule (eq 6) (Duchateau et al. 1995).



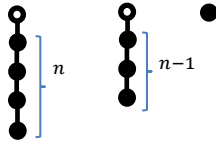
Finally, when the acceptor is the central fructose of a 1-kestose molecule, the reaction produces one glucose and one bifurcose molecule (eq 7). The latter may be further elongated, by 1-FFT or 6-SFT to produce a branched fructan.



The **fructan-fructan 6-G-fructosyltransferase (6G-FFT)**, found in monocots, transfers a fructose unit from a short inulin to the C-6 of the glucose residue of sucrose or inulin, resulting in the formation of a neokestose or inulin neoserie (eq 8). This neokestose or inulin neoserie can be further elongated by 1-FFT or 6-SFT in order to produce inulin or levan neoseries (Ritsema, Joling, and Smeekens 2003).



Two **fructan exohydrolases (FEH)** catalyze the **degradation** of fructans. These enzymes mediate the hydrolysis of the terminal fructose units of a fructan chain (eq 9) which results in the shortening of the fructan molecule and the release of a fructose.



The 1-FEH hydrolyses β (2 $\rightarrow$ 1) glycosidic bonds of inulin and inulo-n-ose while the 6-FEH hydrolyzes β (2 $\rightarrow$ 6) glycosidic bonds of levan and levan-n-ose (W Van den Ende et al. 2001; Wim Van Den Ende et al. 2000; Marx, Nösberger, and Frehner 1997). The 1-FEH and 6-FEH are more common in plant species forming inulin and levan, respectively. However, FEH enzymes may be found in non fructans species (e.g. 6-FEH in sugar beet, Van den Ende, De Coninck et al. 2003). In these species, FEH is involved protection against pathogens by digesting the levan containing slime surrounding some pathogen bacteria (Van Den Ende et al. 2003).

The **acid invertases** complete the breakdown of sucrose into fructose and glucose (eq 10).



2.3.2 Origin of the differences in fructan quality and concentration.

The inter-specific differences in fructan species arise from differences in the enzymes involved in fructan metabolism in each species. The type of fructan largely depends on the type of synthesis enzyme present in a species. The dicot elongation enzymes are 1-SST and 1-FFT. They are involved in inulin and inulo-*n*-ose production. All elongation enzymes mentioned previously are found in monocot, leading to broader fructan spectrum. These substantial differences between monocots and dicots reflect a polyphyletic origin of the fructans enzymes (Vijn and Smeekens 1999). Still, although the same enzymes are present in different species, their activities and substrate affinities drastically influence the mDP of the fructans, as follow.

Inter-specific differences in enzyme **activity** have been reported for 1-FEH. Indeed, a higher activity in *Smallanthus sonchifolius* and *Vernonia herbacea* Vell. Compared to *Viguiera discolor* Baker. would contribute to the higher mDP in *Viguiera discolor* (Itaya et al. 2007).

There are also several illustrations of inter-specific differences of enzyme **affinity** for low or high mDP fructans as fructosyl donor or

acceptor, as reported for 1-FFT, 6-SFT and 1-FEH (Itaya et al. 2007; Tamura et al. 2009; Vergauwen et al. 2003). For example, Vergauwen et al. (2003) compared the chicory 1-FFT to the globe thistle (*Echinops ritro* L.) 1-FFT. The chicory 1-FFT presented a higher affinity for small molecules as fructosyl acceptor such as sucrose, fructose and 1-kestose, resulting in a low mDP. On the other side, globe thistle 1-FFT presented higher affinity for longer fructans as fructosyl acceptor resulting in higher mDP. Similarly, Tamura et al. (2009) observed that the affinity of 1-FFT for 1-kestose as fructosyl acceptor was higher in *V. herbacea* than in the *V. discolor*, resulting in a higher mDP in *V. discolor*. Finally, Itaya et al. (2007) suggested that the high mDP of *V. discolor* was also a consequence of the low affinity of 1-FEH to high DP inulin compared to *S. sonchifolius* and *V. herbacea*. The chicory FEH, however, has equal affinity for low and high DP inulin (Wim Den Van Ende et al. 2000; Wim Van Den Ende et al. 2000). In monocots, the timothy 6-SFT (*Phleum pratense*) has a low affinity for sucrose and a high affinity for long chain fructans resulting in a high mDP (Tamura et al. 2009).

2.3.3 The dynamic of fructans accumulation in chicory is divided in three phases

The dynamics of fructans accumulation in chicory in field from May to January has been divided into three phases by van Arkel et al. (2012), based on different field and controlled experiments and on previous investigations on the carbohydrate metabolism in chicory and several

Asteraceae.

In Belgium, The Netherlands and in the North of France, chicory is generally sown in April and harvested in from September to December. The first phase of fructan accumulation starts around one month after sowing and is characterized by an increase in mDP and in inulin quantity (Figure 5 and Figure 6). The second phase begins in September and is characterized by an increase in inulin quantity and a decrease in mDP. The third phase is triggered by the first night frosts, generally around mid-October. It is characterized by the degradation of inulin, leading to a lower mDP and an increase of the concentration of the undesirable free fructose.

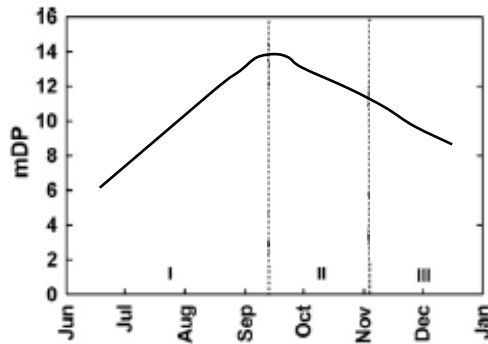


Figure 6: Change in mDP during the cultivation period , inspired from a figure from van Arkel et al. 2012.

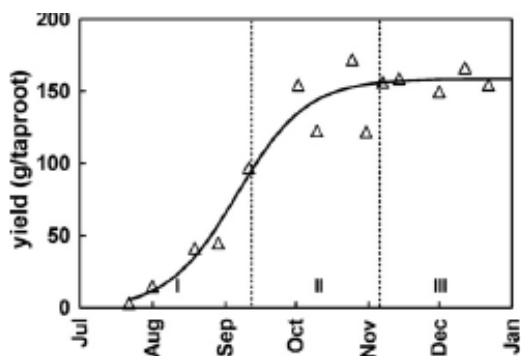
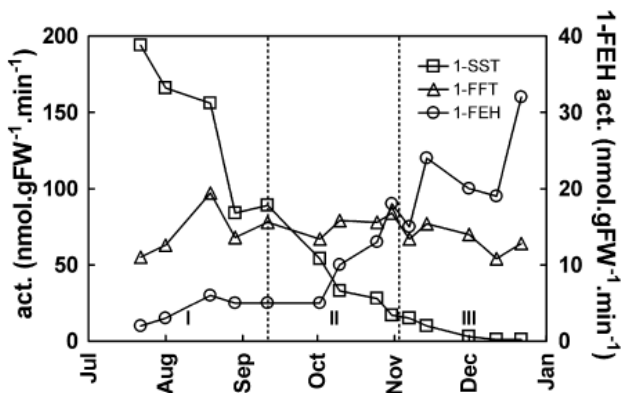


Figure 7: Change in inulin yield of chicory taproot during the cultivation period(van Arkel et al. 2012).

The **first phase** starts at the onset of the radial growth of the root. This phase corresponds to a period of intense **inulin synthesis**, manifested through the increase of the number of inulin molecules, of the mDP and the absence of inulo-*n*-ose. At the end of the first phase, the plant has reached its maximal mDP and has achieved 50 to 70% of its final inulin yield.

The beginning of the first phase coincides with the activation of 1-SST and 1-FFT. The activity of 1-SST reaches its maximum in June and then decreases linearly while that of 1-FFT increases gradually until August and remains constant thereafter. The 1-FEH activity remains very low until the beginning of the second phase (Figure 7).



2

Figure 8: Enzyme activity of the different inulin biosynthesis enzymes. The graph shows enzyme activities in chicory taproot grown in the growing season 2003 of 1-SST (squares), 1-FFT (triangles) and 1-FEH (circles). 1-SST and 1-FFT are expressed as nmol 1-kestose gFW⁻¹ min⁻¹. 1-FEH is expressed as nmol fructose gFW⁻¹ min⁻¹. The dotted vertical lines bound the different phases of the growing season 2003 (van Arkel et al. 2012).

During the first phase, the initiation and elongation of inulin are mediated by the joint activity of 1-SST and 1-FFT. The 1-SST catalyzes the initiation step, by transferring a fructosyl unit from sucrose on another fructose leading to the creation of a free glucose and a 1-Kestose ($GF + GF \rightarrow G + GFF$). The 1-SST activity is higher than the 1-FFT activity, leading to a relatively high 1-Kestose concentration (fig 10). In these conditions, the 1-FFT catalyzes the elongation of inulin molecules by transferring a fructosyl unit from a 1-Kestose to another inulin molecule resulting in sucrose and a longer inulin molecule (eq 2).

For a long time, the decrease in 1-SST activity was suspected to be induced by changes in environmental conditions. However, recent experiments suggest that 1-SST is neither influenced by temperature nor by day length, nor by water deficit. The enzyme appears to be regulated on the post transcriptional level, since its activity correlates with the protein concentration but not with the expression of *1-SST* (Van Laere and Van den Ende 2002).

1-FFT seems to be regulated at the transcriptional level since its activity and the expression of *1-FFT* follow similar patterns. Importantly, this observation has been validated in field conditions and in controlled environments characterized by constant temperature and photoperiod (Dave, Lawrence, and Pennock 2003). Other experiments performed on artificial hairy root culture (HRC) showed that 1-FFT activity increased one week after the *1-FFT* expression. This suggests the existence of other secondary regulation mechanisms of 1-FFT activity (Kusch 2009).

Even though 1-FFT and 1-SST activities follow different patterns, it is worth noticing that the expression of *1-SST* and *1-FFT* are coordinated. Indeed, low nitrogen content and a high fructose/nitrogen (C/N) ratio are inducing the expression of *1-SST* and *1-FFT*, as in barley leaves, where a high C/N ratio induces fructan accumulation (Kusch et al. 2009; Müller et al. 2000). Also, protein kinases (staurosporine) and inhibitors of Ca^{2+} uptake are both inhibiting 1-SST and 1-FFT induction in chicory and other dicots (Kusch et al. 2009; Martínez-Noël et al. 2007).

The second phase generally starts around mid-September and lasts until the first nigh frost in the mid of November. It is characterized by an important inulin synthesis, a substantial increase of the inulin yield and a decrease of mDP. The inulin yield can double during the second phase.

At the beginning of the second phase, the 1-SST activity decreases gradually below that of 1-FFT, resulting in a decrease in glucose, 1-kestose and 1-nystose concentrations (Figure 9). In these conditions, the joint activity of 1-SST and 1-FFT declines and is gradually replaced by the 1-FFT back-transfer activity (Eq 3) which results from the high 1-FFT affinity for sucrose as fructosyl acceptor. While the joint 1-SST and 1-FFT pathway leads to the incorporation of a higher proportion of fructose than glucose in the inulin pool, the back-transfer pathway leads to the incorporation of equal proportion of fructose and glucose in the inulin pool. Therefore, the progressive replacement of the joint 1-SST and 1-FFT activity by the back-transfer pathway leads to a decrease of the mDP inulin and an increase in 1-Kestose and low DP inulin.

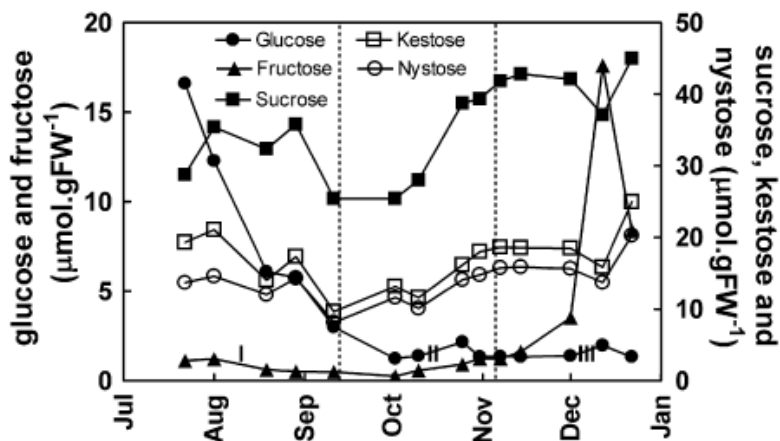


Figure 9: Glucose, fructose, sucrose, 1-kestose and 1-nystose in chicory taproots during the growing season of 2003. Concentration is shown in mol gFW⁻¹ of glucose (closed circles), fructose (closed triangles), sucrose (closed squares), 1-kestose (open squares) and 1-nystose (open circles) in chicory taproot (van Arkel et al. 2012).

Some 1-FEH activity is initiated during the second phase, at the beginning of October. The 1-FEH activity contributes to inulin breakdown and therefore to the decrease of the mDP. However, its activity remains very low during the second phase and does not cause any significant increase in free fructose content.

The third phase occurs when the first frost initiates the degradation of inulin and the reduction of the tap root growth. The degradation of inulin results in the decrease of mDP, an increase of free fructose, gives rise to inulose

synthesis, and the stabilization of the inulin yield.

The third phase is characterized by a strong increase in 1-FEH activity. Three FEH are known in chicory, *i.e.* 1-FEH I, 1-FEH IIa and 1-FEH IIb. Due to their high similarities, *1-FEH IIa* and *1-FEH IIb* expressions cannot be distinguished. It is now clear that *1-FEH I* and *1-FEH II* expressions are both directly initiated by low temperatures and that their protein activities are negatively correlated to temperature (Maroufi et al. 2012). Nevertheless, 1-FEH I and 1-FEH II have different temperature set points. *1-FEH I* is expressed when temperature goes below 10°C, and is followed by *1-FEH II* after the first frost (or < 4 °C, according to the experiments) (Figure 10). Not surprisingly, the persistence of 1-FEH I after cold spell at 4 °C (11 days) is superior to that of 1-FEH II (4 days). The temperature response of 1-FEH genes depends also on the genotype. When roots from different cultivars are stored at 4°C for 9 weeks, some cultivars show a constant or increasing expression of *1-FEH I* while *1-FEH II* is always increasing during this period.

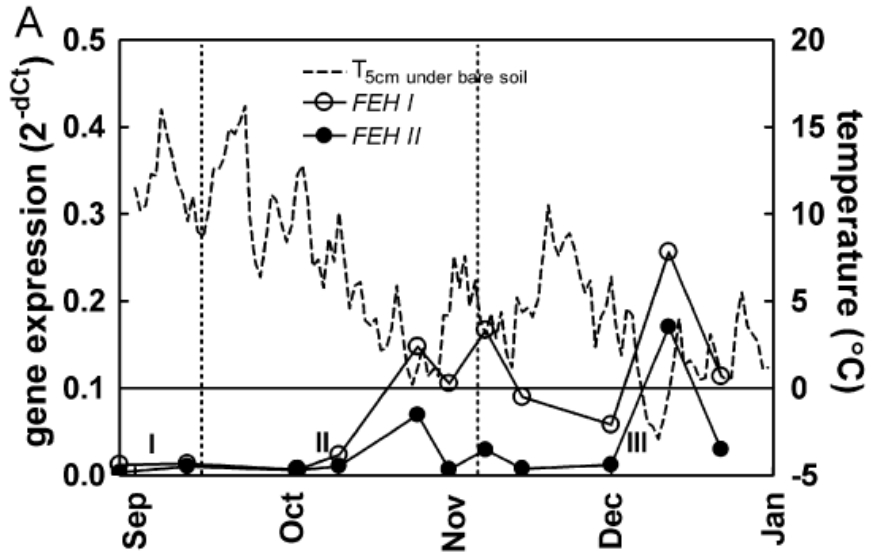


Figure 10: Relative gene expression of 1-FEH I and 1-FEH II in field-grown taproots and the minimum day temperature, during the growing season. The relative expression of 1-FEH I and 1-FEH II was measured in taproots grown during the growing season and plotted on the left axis. The minimum day temperature 5 cm under bare soil (striped line) is plotted on the right y-axis. The dotted vertical lines bound the different phases (van Arkel et al. 2012).

The activities of the different 1-FEHs are also affected by other factors. *1-FEH I* and *1-FEH IIa* expressions are differentially affected by inhibitors such as protein kinases, protein phosphatases and Ca^{2+} (Kusch et al. 2009; Michiels et al. 2004). *1-FEH II* is also inhibited by sucrose while *1-FEH I* is not (Claessens, Van Laere, and De Proft 1990; Wim Den Van Ende et al. 2000; De Roover, Van Laere, and Van den Ende 1999; Verhaest et al. 2007).

Therefore, during the third phase, the 1-FEH activity and expression increase, the 1-SST activity is extremely low and the 1-FFT back-transfer activity remains constant. In addition, 1-FFT binds substantial fraction of the free fructose released by the 1-FEH activity with other fructose molecules, resulting in the formation of inulose (eq 4). The resulting effect is a decrease of mDP, a release of free fructose and inulose, at constant inulin concentration.

2.4 Fructan profile in chicory: an ideotype

According to the informations reported in the preceding sections, we foresee three ideotypes of chicory that exploit changes in expression, activity or affinity of the enzymes of the fructan pathway in order to improve inulin yield and quality.

The first ideotype would have an **extended duration of 1-SST expression** leading to a longer first phase (Figure 11). The 1-kestose offer would be maintained over a longer period, which would delay the 1-FFT back-transfer activity. Such ideotype would achieve a **higher mDP** at constant inulin yield and fructose concentration. As a proof of concept, two transgenic chicory lines containing the 1-SST I or 1-SST II gene from Jerusalem artichoke presented an up to 20 % higher mDP than control line (Van Arkel et al. 2013).

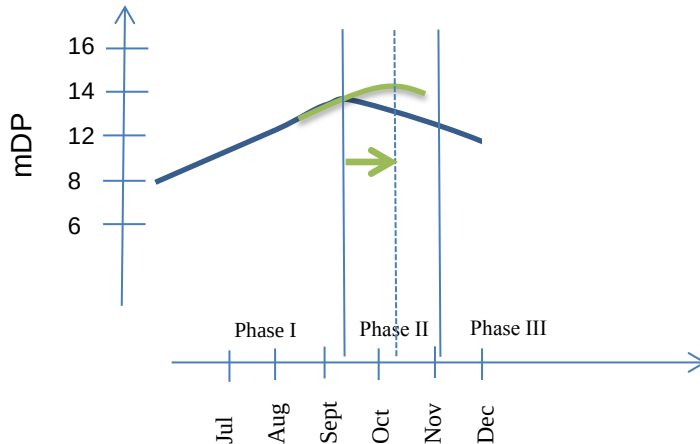


Figure 11: Change in inulin mDP in normal chicory taproot (dark blue, inspired from a figure from van Arkel et al. 2012.) and the chicory ideotype presenting an extended duration of 1-SST expression (green).

The second ideotype would have a **higher affinity of 1-FFT for long or short fructosyl donor and acceptor** (Figure 12), leading to an increase in the mDP and a delay of the onset of back-transfer activity, at constant inulin yield, fructose concentration and phase duration. In particular, a higher affinity for higher DP fructans as fructosyl acceptor could increase the mDP and decrease the inulose synthesis. A higher affinity for low DP molecule, 1-kestose, and sucrose could reduce the mDP, increase the inulose synthesis without increasing the free fructose content.

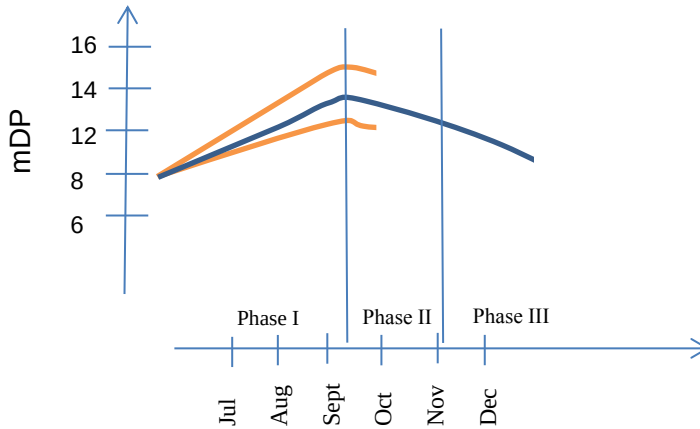


Figure 12: Change in inulin mDP in normal chicory taproot (dark blue, inspired from a figure from van Arkel et al. 2012.) and the chicory ideotype characterized by high or low 1-FFT affinity for low DP inulin (orange).

The third ideotype would have a **reduced or delayed 1-FEH activity**, leading to higher mDP and inulin yield with lower fructose and inulo-*n*-ose content (Figure 13). A first attempt to decrease the 1-FEH activity in chicory line expressing 1-FEH I antisense was realized by Van Arkel (2013). However, even after cold treatment, the transgenic line exhibited only few differences compared to control lines.

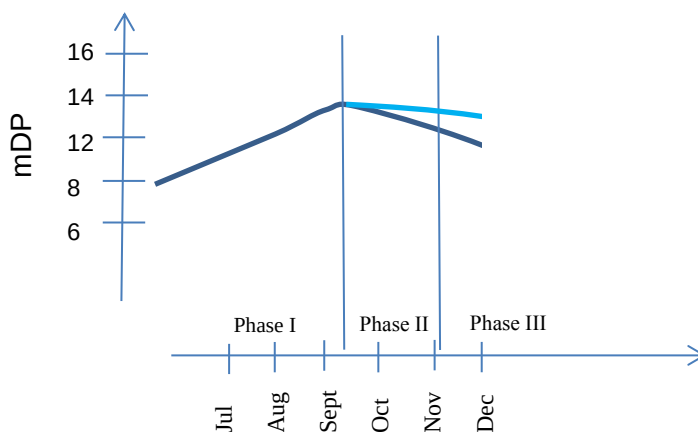


Figure 13: Change in inulin mDP in normal chicory taproot (dark blue, inspired from a figure from van Arkel et al. 2012.) and in the chicory ideotype with a low 1-FEH activity.

2.5 Genetic determinism of fructans

2.5.1 QTL involved in fructan quality has been identified in different species.

Within fructan-producing species, the concentration and the type of fructan can vary across individuals which reveal an intra-specific phenotypic diversity. Different studies were undertaken to identify genetic factors underlying this phenotypic variation. In particular, QTL

and candidate gene association analysis were conducted in wheat, onion and ryegrass.

In wheat, QTLs affecting the fructan content were identified on the chromosomes 2B, 3B, 5A, 6D and 7A (Huynh et al. 2008, 2012; Zhang et al. 2008a). The QTL on chromosomes 6D and 7A had the largest effect and accounted for 17% and 27% of the total phenotypic variance, respectively. Interestingly, the QTL located on the chromosome 7A overlaps a cluster of genes encoding for 1-SST, 1-FFT and 6-SFT. This cluster is under the control of an unique transcriptional activator (TaMYB13) and is also present in the barley genome (*Hordeum vulgare*)(Huynh et al. 2012). The QTL located on chromosome 6D correspond to the 1-FEH location (Zhang et al. 2008b).

In onion, a major-effect QTL affecting the concentration of fructans and possibly other carbohydrates (fructose, glucose and sucrose) was identified on chromosomes 8 associated to 93 % of the phenotypic variation in fructan content in a bi-parental segregation population (McCallum et al. 2006; Raines, Henson, and Havey 2009; Yaguchi et al. 2008) . The candidate gene associated to this QTL is a sucrose phosphate synthase (SPS) and may have a key role in conditioning the sucrose level. In some species, 1-SST expression is controlled by the sucrose level(Müller et al. 2000; Wang, Van den Ende, and Tillberg 2000).

In ryegrass (*Lolium perenne*), a QTL analysis identified 3 QTLs explaining between 8 and 59 % of the phenotypic variability for different fructans and carbohydrate-related parameters (Turner et al. 2006). Later, a candidate gene-based association study revealed the association between fructans content and a polymorphism in a gene functionally associated with stem elongation (LpCAD2). This polymorphism was associated with 8 % of the phenotypic variability. The underlying mechanism is still unknown but the authors suspect that this gene may affect carbon partitioning in the plant (Pembleton et al. 2013). Some genes involved in the carbohydrate metabolism pathway were also correlated with genes involved in fiber content. Again, the underlying mechanisms are unknown.

The existence of QTL intervals lacking obvious candidate genes and the evidence of correlations between fructans and various aspects of carbon metabolism suggest a complex genetic determinism involving pleiotropic effects or transcription factors affecting the expression of genes involved in carbohydrates (including fructans) metabolism.

2.5.2 Chicory presents exploitable genetic variability

As reported above, the major genes involved in fructan metabolism in chicory have been identified and their expression during the season has

been determined. Surprisingly, little is known on the genetic variability of fructan metabolism in chicory. There are, however, a number of indications that chicory presents much genetic variability that can be exploited in breeding.

We have shown that the fructan profile and yield of carbohydrate largely depends on crop management, i.e. sowing and harvest dates and post-harvest storage duration (Amaducci and Pritoni 1998; Baert 1997; Ernst, Chatterton, and Harrison 1995; Wilson, Smith, and Yonts 2004). However, important phenotypic differences of fructan profile and yield also exist between lines and cultivars (Amaducci and Pritoni 1998; Baert 1997; Wilson, Smith, and Yonts 2004 Koch et al. 1999). Some of these differences arise from the restrictions of crop management imposed by specific cultivars. Indeed, farmers try to sow the chicory as early as possible in order to increase the cultivation duration and the yield. However, early sowing faces two important limitations. Firstly, cold spring conditions lead to the bolting of chicory at the end of the first season, resulting in a lower root quality and increased difficulty of carbohydrate extraction. Moreover, the seeds produced following bolting become weeds for the next cropping season. Secondly, cold temperatures hampers the growth rate of the plant (Venema, Eekhof, and Hasselt 2000; Wolfe 1991) and, if combined with high irradiance, reduce the photosynthetic efficiency (Alves, Magalhães, and Barja 2002). On this respect, different cultivars present different susceptibilities to these limiting spring conditions (Devacht et al. 2009). Kosh et al. (1999) also reported cultivar differences for the yield and mDP at harvest after cold exposure.

Others physiological parameters also present some genetic differences, as indicated by their differences among varieties. These include leaf stand, leaf area, bolting percentage, root yield, total sugar content and the degree of fructan polymerization (Wilson, Smith, and Yonts 2004).

Few investigations of the genetic bases of trait in chicory were implemented in chicory. Among them, one candidate-gene study has been undertaken to identify polymorphism underlying carbohydrate quality variation in chicory. This unique study describes differential expression of 1-FEH IIa among Witloof and two industrial chicory varieties after a period of storage (Maroufi, Van Bockstaele, and De Loose 2012). An other one consisted in the identification of QTLs associated to pith characteristics, taste, yield, sensitivity to nitrogen fertilization and the cartography of male sterility factors in a F2 populations, issued from two Witloof individuals (Cassan et al. 2010; Gonthier et al. 2013; Van Stallen et al. 2005a, 2005b). The creation of segregation population, required for QTL analysis is complicated by the the bisexual nature of chicory flower and the short duration of flower opening (few hours at dawn) make the crossing process complex and time consuming. Moreover, the partial incompatibility system of chicory makes the crossing results very uncertain.

2.6 Reference

- Alves, P.L., Magalhães, A.C.N., & Barja, P.R.. (2002). The Phenomenon of Photoinhibition of Photosynthesis and Its Importance in Reforestation. *The Botanical Review* 68: 193–208.
- Amaducci, S., & Pritoni, G. (1998). Effect of Harvest Date and Cultivar on *Cichorium Intybus* Yield Components in North Italy. *Industrial Crops and Products* 7: 345–49.
- Baert, J. R. (1997). The effect of sowing and harvest date and cultivar on inulin yield and composition of chicory (*Cichorium intybus* L.) roots. *Industrial Crops and Products*, 6(3), 195-199.
- Beirão-da-Costa, M. L., Januário, M. I. N., Simão, F. M. S., & leitão, A. E. B. (2009). Characterisation of inulin from chicory and salsify cultivated in Portugal. *Alimentos e Nutrição Araraquara*, 16(3), 221-225.
- Cadalen, T., Mörchen, M., Blassiau, C., Clabaut, A., Scheer, I., Hilbert, J. L., ... & Quillet, M. C. (2010). Development of SSR markers and construction of a consensus genetic map for chicory (*Cichorium intybus* L.). *Molecular Breeding*, 25(4), 699-722.
- Cassan, L., Moreau, L., Segouin, S., Bellamy, A., Falque, M., & Limami, A. M. (2010). Genetic map construction and quantitative trait loci (QTL) mapping for nitrogen use efficiency and its relationship with productivity and quality of the biennial crop Belgian endive (*Cichorium intybus* L.). *Journal of plant Physiology*, 167(15), 1253-1263.
- Chubey, B. B., & Dorrell, D. G. (1978). Total reducing sugar, fructose and glucose concentrations and root yield of two chicory cultivars as affected by

- irrigation, fertilizer and harvest dates. *Canadian Journal of Plant science*, 58(3), 789-793.
- Claessens, G., Van Laere, A., & De Proft, M. (1990). Purification and properties of an inulinase from chicory roots (*Cichorium intybus* L.). *Journal of Plant Physiology*, 136(1), 35-39.
- Devacht, S., Lootens, P., Roldan-Ruiz, I., Carlier, L., Baert, J., Van Waes, J., & Van Bockstaele, E. (2009). Influence of low temperatures on the growth and photosynthetic activity of industrial chicory, *Cichorium intybus* L. partim. *Photosynthetica*, 47(3), 372-380.
- Duchateau, N., Bortlik, K., Simmen, U., Wiemken, A., & Bancal, P. (1995). Sucrose: fructan 6-fructosyltransferase, a key enzyme for diverting carbon from sucrose to fructan in barley leaves. *Plant Physiology*, 107(4), 1249-1255.
- Edelman, J., & Jefford, T. G. (1968). The mechanism of fructosan metabolism in higher plants as exemplified in *Helianthus tuberosus*. *New phytologist*, 517-531.
- Eenink, A. H. (1981). Compatibility and incompatibility in witloof-chicory (*Cichorium intybus* L.). 2. The incompatibility system. *Euphytica*, 30(1), 77-85.
- Van den Ende, W., Michiels, A., Van Wouterghem, D., Clerens, S. P., De Roover, J., & Van Laere, A. J. (2001). Defoliation induces fructan 1-exohydrolase II in witloof chicory roots. Cloning and purification of two isoforms, fructan 1-exohydrolase IIa and fructan 1-exohydrolase IIb. Mass fingerprint of the fructan 1-exohydrolase II enzymes. *Plant Physiology*, 126(3), 1186-1195.
- Van den Ende, W., Michiels, A., De Roover, J., Verhaert, P., & Van Laere, A.

(2000). Cloning and functional analysis of chicory root fructan 1-exohydrolase I (1-FEH I): a vacuolar enzyme derived from a cell-wall invertase ancestor? Mass fingerprint of the 1-FEH I enzyme. *The Plant Journal*, 24(4), 447-456.

Van den Ende, W., Michiels, A., Van Wouterghem, D., Vergauwen, R., & Van Laere, A. (2000). Cloning, developmental, and tissue-specific expression of sucrose: sucrose 1-fructosyl transferase from *Taraxacum officinale*. Fructan localization in roots. *Plant Physiology*, 123(1), 71-80.

Van den Ende, W., De Coninck, B., Clerens, S., Vergauwen, R., & Van Laere, A. (2003). Unexpected presence of fructan 6-exohydrolases (6-FEHs) in non-fructan plants: characterization, cloning, mass mapping and functional analysis of a novel 'cell-wall invertase-like' specific 6-FEH from sugar beet (*Beta vulgaris* L.). *The Plant Journal*, 36(5), 697-710.

Van den Ende, W., & Van Laere, A. (1996a). De-novo synthesis of fructans from sucrose in vitro by a combination of two purified enzymes (sucrose: sucrose 1-fructosyl transferase and fructan: fructan 1-fructosyl transferase) from chicory roots (*Cichorium intybus* L.). *Planta*, 200(3), 335-342.

Van den Ende, W., & Van Laere, A. (1996b). Fructan synthesizing and degrading activities in chicory roots (*Cichorium intybus* L.) during field-growth, storage and forcing. *Journal of Plant Physiology*, 149(1), 43-50.

Ernst, M., Chatterton, N. J., & Harrison, P. A. (1995). Carbohydrate changes in chicory (*Cichorium intybus* L. var. *foliosum*) during growth and storage. *Scientia Horticulturae*, 63(3), 251-261.

Esau, K. (1965). Plant anatomy. 2nd edn. *Wiley International Edition*.

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, IV. (2003). Structure of linkage disequilibrium in plants*. *Annual review of plant biology*, 54(1), 357-

374.

Gonthier, L., Blassiau, C., Mörchen, M., Cadalen, T., Poiret, M., Hendriks, T., & Quillet, M. C. (2013). High-density genetic maps for loci involved in nuclear male sterility (NMS1) and sporophytic self-incompatibility (S-locus) in chicory (*Cichorium intybus* L., Asteraceae). *Theoretical and Applied Genetics*, 126(8), 2103-2121.

Hellwege, E. M., Raap, M., Gritscher, D., Willmitzer, L., & Heyer, A. G. (1998). Differences in chain length distribution of inulin from *Cynara scolymus* and *Helianthus tuberosus* are reflected in a transient plant expression system using the respective 1-FFT cDNAs. *FEBS letters*, 427(1), 25-28.

Hendry, G. (1993). Evolutionary of and and Appraisal Natural Functions Fructans a Climatological , *Biogeographic Mechanistic*. 123(1): 3–14.

Huynh, B. L., Wallwork, H., Stangoulis, J. C., Graham, R. D., Willsmore, K. L., Olson, S., & Mather, D. E. (2008). Quantitative trait loci for grain fructan concentration in wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics*, 117(5), 701-709.

Huynh, B. L., Mather, D. E., Schreiber, A. W., Toubia, J., Baumann, U., Shoaie, Z., ... & Fleury, D. (2012). Clusters of genes encoding fructan biosynthesizing enzymes in wheat and barley. *Plant molecular biology*, 80(3), 299-314.

Itaya, N. M., Asega, A. F., Carvalho, M. A. M., & Rita de Cássia, L. (2007). Hydrolase and fructosyltransferase activities implicated in the accumulation of different chain size fructans in three Asteraceae species. *Plant Physiology and Biochemistry*, 45(9), 647-656.

Kiers, A. M., Mes, T. H., Meijden, R. V. D., & Bachmann, K. (2000). A

search for diagnostic AFLP markers in *Cichorium* species with emphasis on endive and chicory cultivar groups. *Genome*, 43(3), 470-476.

Koch, K., Andersson, R., Rydberg, I., & Aman, P. (1999). Influence of harvest date on inulin chain length distribution and sugar profile for six chicory. *Journal of the Science of Food and Agriculture*, 79, 1503-1506.

Koopman, W., (2002). Zooming in on the lettuce genome: species relationships in *Lactuca* s.l., inferred from chromosomal and molecular characters. *PhD Thesis*, Wageningen University, Netherlands

Kusch, U., Harms, K., Rausch, T., & Greiner, S. (2009). Inhibitors of plant invertases do not affect the structurally related enzymes of fructan metabolism. *New Phytologist*, 181(3), 601-612.

Van Laere, A., & Van den Ende, W. (2002). Inulin metabolism in dicots: chicory as a model system. *Plant, Cell & Environment*, 25(6), 803-813.

Livingston, D. P., Chatterton, N. J., & Harrison, P. A. (1993). Structure and quantity of fructan oligomers in oat (*Avena* spp.). *New Phytologist*, 123(4), 725-734.

López-Molina, D., Navarro-Martínez, M. D., Rojas-Melgarejo, F., Hiner, A. N., Chazarra, S., & Rodríguez-López, J. N. (2005). Molecular properties and prebiotic effect of inulin obtained from artichoke (*Cynara scolymus* L.). *Phytochemistry*, 66(12), 1476-1484.

Maroufi, A., Van Bockstaele, E., De Loose M.. 2012. "Differential Expression of Fructan 1-Exohydrolase Genes Involved in Inulin Biodegradation in Chicory (*Cichorium Intybus*) Cultivars." *Australian Journal of Crop Science* 6(9): 1362-68.

Martínez-Noël, G., Nagaraj, V. J., Caló, G., Wiemken, A., & Pontis, H. G. (2007). Sucrose regulated expression of a Ca²⁺-dependent protein kinase

(TaCDPK1) gene in excised leaves of wheat. *Plant Physiology and Biochemistry*, 45(6), 410-419.

Marx, S. P., JOSEF, N., & Frehner, M. (1997). Hydrolysis of fructan in grasses: a β -(2-6)-linkage specific fructan- β -fructosidase from stubble of *Lolium perenne*. *New Phytologist*, 135(2), 279-290.

McCallum, J., Clarke, A., Pither-Joyce, M., Shaw, M., Butler, R., Brash, D., ... & Havey, M. J. (2006). Genetic mapping of a major gene affecting onion bulb fructan content. *Theoretical and applied genetics*, 112(5), 958-967.

Michiels, A., Van Laere, A., Van den Ende, W., & Tucker, M. (2004). Expression analysis of a chicory fructan 1-exohydrolase gene reveals complex regulation by cold. *Journal of experimental botany*, 55(401), 1325-1333.

Müller, J., Aeschbacher, R. A., Sprenger, N., Boller, T., & Wiemken, A. (2000). Disaccharide-mediated regulation of sucrose: fructan-6-fructosyltransferase, a key enzyme of fructan synthesis in barley leaves. *Plant Physiology*, 123(1), 265-274.

Muys, C., Thienpont, C. N., Dauchot, N., Maudoux, O., Draye, X., & Cutsem, P. V. (2014). Integration of AFLPs, SSRs and SNPs markers into a new genetic map of industrial chicory (*Cichorium intybus* L. var. sativum). *Plant breeding*, 133(1), 130-137.

Nguyen-Quoc, B., & Foyer, C. H. (2001). A role for 'futile cycles' involving invertase and sucrose synthase in sucrose metabolism of tomato fruit. *Journal of Experimental Botany*, 52(358), 881-889.

Pembleton, L. W., Wang, J., Cogan, N. O. I., Pryce, J. E., Ye, G., Bandaranayake, C. K., ... & Forster, J. W. (2013). Candidate gene-based association genetics analysis of herbage quality traits in perennial ryegrass

(*Lolium perenne* L.). *Crop and Pasture Science*, 64(3), 244-253.

Pilon-Smits, E., Ebskamp, M., Paul, M.J., Jeuken, M., Weisbeek, P.J., Smeeckens, S. (1995) Improved Performance of Transgenic Fructan-Accumulating Tobacco under Drought Stress. *Plant Physiology*, 107(1):125-130.

Pollock, C. J. (1986). Tansley Review No. 5. Fructans and the metabolism of sucrose in vascular plants. *New Phytologist*, 1-24.

Raines, S., Henson, C., & Havey, M. J. (2009). Genetic analyses of soluble carbohydrate concentrations in onion bulbs. *Journal of the American Society for Horticultural Science*, 134(6), 618-623.

Ritsema, T., Joling, J., & Smeeckens, S. (2003). Patterns of fructan synthesized by onion fructan: fructan 6G-fructosyltransferase expressed in tobacco BY2 cells—is fructan: fructan 1-fructosyltransferase needed in onion?. *New Phytologist*, 160(1), 61-67.

De Roover, J., Van Laere, A., & Van den Ende, W. (1999). Effect of defoliation on fructan pattern and fructan metabolizing enzymes in young chicory plants (*Cichorium intybus*). *Physiologia Plantarum*, 106(2), 158-163.

Shiomi, N. (1989). Properties of fructosyltransferases involved in the synthesis of fructan in liliaceous plants. *Journal of Plant Physiology*, 134(2), 151-155.

De Simone, M., Morgante, M., Lucchin, M., Parrini, P., & Marocco, A. (1997). A first linkage map of *Cichorium intybus* L. using a one-way pseudotestcross and PCR-derived markers. *Molecular Breeding*, 3(6), 415-425.

Stallen, N. V., Vandenbussche, B., Londers, E., Noten, V., & Proft, M. D. (2005a). QTL analysis of important pith characteristics in a cross between two inbred lines of chicory (*Cichorium intybus* var. *foliosum*): a preliminary

study. *Plant breeding*, 124(1), 54-58.

Stallen, N. V., Vandenbussche, B., Londers, E., Noten, V., & Proft, M. D. (2005b). QTL analysis of production and taste characteristics of chicory (*Cichorium intybus* var. *foliosum*). *Plant breeding*, 124(1), 59-62.

Stallen, N. V., Vandenbussche, B., Verdoodt, V., & Proft, M. D. (2003). Construction of a genetic linkage map for witloof (*Cichorium intybus* L. var. *foliosum* Hegi). *Plant breeding*, 122(6), 521-525.

Tamura, K. I., Kawakami, A., Sanada, Y., Tase, K., Komatsu, T., & Yoshida, M. (2009). Cloning and functional analysis of a fructosyltransferase cDNA for synthesis of highly polymerized levans in timothy (*Phleum pratense* L.). *Journal of experimental botany*, 60(3), 893-905.

Turner, L. B., Cairns, A. J., Armstead, I. P., Ashton, J., Skøt, K., Whittaker, D., & Humphreys, M. O. (2006). Dissecting the regulation of fructan metabolism in perennial ryegrass (*Lolium perenne*) with quantitative trait locus mapping. *New Phytologist*, 169(1), 45-58.

van Arkel, J., Vergauwen, R., Sévenier, R., Hakkert, J.C., van Laere, A., Bouwmeester, H. J., ... & van der Meer, I.M. (2012) Sink Filling, Inulin Metabolizing Enzymes and Carbohydrate Status in Field Grown Chicory (*Cichorium Intybus* L.).” *Journal of Plant Physiology* 169:1520–29.

van Arkel, J. , Sévenier, R. , Hakkert, J. C. , Bouwmeester, H. J., Koops, A. J., & van der Meer, I. M. (2013). Tailor-made fructan synthesis in plants: a review. *Carbohydrate Polymers*, 93(1), 48-56.

Van den Ende, W., De Coninck, B., Clerens, S., Vergauwen, R., & Van Laere, A. (2003). Unexpected presence of fructan 6-exohydrolases (6-FEHs) in non fructan plants: characterization, cloning, mass mapping and functional

analysis of a novel 'cell-wall invertases like' specific 6-FEH from sugar beet (*Beta vulgaris* L.). *The Plant Journal*, 36(5), 697-710.

Venema, J. H., Eekhof, M., & van Hasselt, P. R. (2000). Analysis of low-temperature tolerance of a tomato (*Lycopersicon esculentum*) cybrid with chloroplasts from a more chilling-tolerant *L. hirsutum* accession. *Annals of Botany*, 85(6), 799-807.

Vergauwen, R., Van Laere, A., & Van den Ende, W. (2003). Properties of fructan: fructan 1-fructosyltransferases from chicory and globe thistle, two Asteracean plants storing greatly different types of inulin. *Plant physiology*, 133(1), 391-401.

Verhaest, M., Lammens, W., Le Roy, K., De Ranter, C. J., Van Laere, A., Rabijns, A., & Van den Ende, W. (2007). Insights into the fine architecture of the active site of chicory fructan 1-exohydrolase: 1-kestose as substrate vs sucrose as inhibitor. *New Phytologist*, 174(1), 90-100.

Vijn, I., & Smeekens, S. (1999). Fructan: more than a reserve carbohydrate?. *Plant physiology*, 120(2), 351-360.

Wang, C., Van den Ende, W., & Tillberg, J. E. (2000). Fructan accumulation induced by nitrogen deficiency in barley leaves correlates with the level of sucrose: fructan 6-fructosyltransferase mRNA. *Planta*, 211(5), 701-707.

Wilson, R. G., Smith, J. A., & Yonts, C. D. (2004). Chicory root yield and carbohydrate composition is influenced by cultivar selection, planting, and harvest date. *Crop Science*, 44(3), 748-752.

Wieser F., (1965). Am5er F., 2. yield and carbohydrate composition is infrealization des objectifs indispensables à cette fin, dans l'optique de la mécanisation totale de cette culture. *Annales de Jembloux* Vol. 71, 177-186.

Wolfe, D. W. (1991). Low temperature effects on early vegetative growth, leaf gas exchange and water potential of chilling-sensitive and chilling-tolerant crop species. *Annals of Botany*, 67(3), 205-212.

Yaguchi, S., McCallum, J., Shaw, M., Pither-Joyce, M., Onodera, S., Shiomi, N., ... & Shigyo, M. (2008). Biochemical and genetic analysis of carbohydrate accumulation in *Allium cepa* L. *Plant and cell physiology*, 49(5), 730-739.

Zhang, J., Huang, S., Fosu-Nyarko, J., Dell, B., McNeil, M., Waters, I., ... & Appels, R. (2008a). The genome structure of the 1-FEH genes in wheat (*Triticum aestivum* L.): new markers to track stem carbohydrates and grain filling QTLs in breeding. *Molecular Breeding*, 22(3), 339-351.

Zhang, J., Huang, S., Fosu-Nyarko, J., Dell, B., McNeil, M., Waters, I., ... & Appels, R. (2008b). The genome structure of the 1-FEH genes in wheat (*Triticum aestivum* L.): new markers to track stem carbohydrates and grain filling QTLs in breeding. *Molecular Breeding*, 22(3), 339-351.

3 An introduction to population and quantitative Genetics.

Breeders exploit the genetic diversity by generating new individuals and selecting the best ones based on their phenotypic performances. This selection process aims at increasing the frequency of favorable alleles and allele combinations for several traits (for example in chicory: inulin quality, root shape, disease resistance...). Recently, marker-assisted selection (MAS) has empowered breeding strategies by enabling the direct selection of favorable alleles. This approach requires *a priori* knowledge of the genetic factors underlying heritable phenotypic variation.

A prerequisite to any work aiming at understanding the genetic base of the variation of a trait consist at evaluating the available phenotypic and genetic variation and their distribution within the population. Therefore, the first section of this chapter investigates three aspects of the genetic diversity: (i) The genetic diversity, (ii) the distribution of the genetic diversity of in a population (population structure) and finally, (iii), the relation that may occur between the genetic diversity present in different loci (linkage disequilibrium).

The second section of this chapter aims at describing the methods developed to determine the genetic factors at the origin of the variation of a trait. Therefore, the first part of this section describes the different components of the determinism of a trait. The second part provides an overview of quantitative trait locus (QTL) mapping and genome wide association studies (GWAS), two methods developed to identify genetic factors involved into the determinism of a trait. The third part details some theoretical and technical aspects of the GWAS.

3.1 Population genetics

3.1.1 Genetic diversity

Diversity is the degree of variation of the elements present in a system. In the case of genetic diversity, the elements are DNA polymorphisms. Therefore a usual approach of the genetic diversity consists in the evaluation of the number of polymorphic loci or the number of alleles present in multiallelic markers such as SNPs, SSR or AFLP or into the density of polymorphic loci identified in a DNA sequence. The second approach integrates the variance aspect of the diversity. The variance of a system is maximized when all the elements are equally represented -in the case of genetic markers, when the different alleles are present the same frequencies-. The Nei genetic index (I_{Nei}) is a measure of the genetic diversity that integrates the variance component (Nei, 1973). Considering a biallelic locus with two alleles A and a, the Nei genetic index is:

$$I_{\text{Nei}} = 1 - (p_A^2 + p_a^2)$$

p_A is the frequency of the A allele

p_a is the frequency of the a allele

When it is extended to different multiallelic loci:

$$I_{\text{Nei}} = 1 - \frac{(\sum \sum p_{ij}^2)}{n}$$

p_{ij} is the frequency of the i^{th} allele of the j^{th} locus

n is the number of loci

I_{Nei} ranges from 0 to 1, increases with the number of alleles, and for a specific allele number, is maximum when all alleles are equally represented, i.e., have the same frequencies.

The genotype is a term used to describe the allele combination present in one individual, for one or several loci. In the present section it refers to a single locus. The Hardy-Weinberg principle describes the relation between the genotypes and alleles frequencies in an ideal population (also referred to as random mating or Hardy-Weinberg population) defined by random mating (allogamy, no structure, no migration, no selection), infinite size, non-overlapping generation and no mutation (Weinberg, 1908 and Hardy 1908). In such population, the alleles present in each genotype are randomly sampled in the allele pool of the previous generation and the allelic frequencies are stable from one generation to one other. Thus the genotype frequencies directly depend on allele frequencies and may be evaluated using a contingency table (Table 1). In a diploid species, in the particular case of a bi-allelic locus presenting the alleles A and a with frequencies p_A and p_a respectively, the homozygotes AA and aa frequencies are p_A^2 and p_a^2 respectively and the heterozygote frequency is $2 p_A p_a$ (Figure 14). Note that

as an allele becomes rarer, the frequency of homozygous genotype for this allele is much lower than the frequency of the heterozygous genotype.

		Male gametes	
		A (p_A)	a (p_a)
Female gametes	A (p_A)	AA (p_A^2)	Aa ($p_A p_a$)
	a (p_a)	Aa ($p_a p_A$)	aa (p_a^2)

Table 1: Contingency table showing allele and genotype frequencies of a bi-allelic locus in an in Hardy-Weinberg population. The frequencies of the alleles A and a are p_A and p_a respectively; the resulting frequencies of the genotypes AA, aa and Aa are p_A^2 , p_a^2 and $2p_A p_a$.

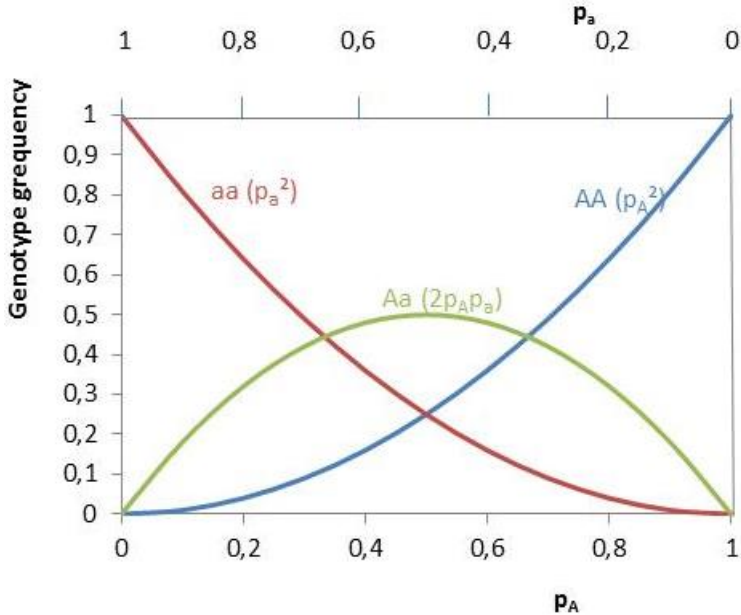


Figure 14: Evolution of the frequencies of the genotypes AA, Aa and aa genotypes frequencies according to the allele frequencies A and a with frequencies p_A and p_a respectively of a biallelic locus of a diploid population in Hardy-Weinberg equilibrium.

The expected heterozygote frequency under Hardy-Weinberg condition (H_e) corresponds to I_{Nei} . Therefore, for a bi-allelic locus of a diploid population:

$$H_e = I_{Nei} = 2p_A p_a = 1 - (p_A^2 + p_a^2)$$

When it is extended to several multi allelic loci:

$$H_e = I_{Nei} = 1 - \frac{(\sum \sum p_{ij}^2)}{n}$$

p_{ij} is the frequency of the i th allele of the j th locus

n is the number of loci

Any disturbance of the conditions defining the Hardy-Weinberg equilibrium results in a deviation of the observed genotype frequencies from their expectation. Random mating, for example may be broken by the population structure, autogamy, migration, selection or most of other evolutionary forces.

3.1.2 Population structure

Population structure refers to the distribution of the genetic diversity among the individuals of a population. In particular, it is characterized by the occurrence of groups of individuals (sub-populations) presenting allele frequencies that differ from each other.

The population structure arises when groups of individuals are isolated from each other. There, the sub-population independently experiment processes that affect their allele frequencies. These processes are divided in to two classes: the systematic and the dispersive processes. The systematic processes tend to change the allele frequencies in a predictable manner both in amount and in direction. They include mutation, migration and selection. The mutation consists in the apparition of new allele and mainly depends on the

mutation rate. The migration consists in the introduction, in a population of new individuals presenting other allelic frequencies. The resulting allele frequencies of the population depend on the rate of migration and the allele frequencies present in the migrant individuals. Finally, the selection is the fact that, in a particular environment, some traits make it more likely for an organism to survive and reproduce. It results in the increase in frequencies of the alleles at the origin of advantageous trait. The importance of the change in allele frequencies is proportional to the selection pressure. The dispersive processes arise in small populations from the effects of random sampling of different alleles from one generation from another (Figure 15). These changes are predictable in amount but not in direction. This process result in the genetic drift: i.e. the random change of allele frequencies from one generation to one other. Extreme genetic drift can lead to the disappearance of an allele in a population and therefore, to the genetic uniformity within the population (Thiagarajan, 2014, Acquaa, 2012, Hartl et al. 2007).

An introduction to population and quantitative Genetics.

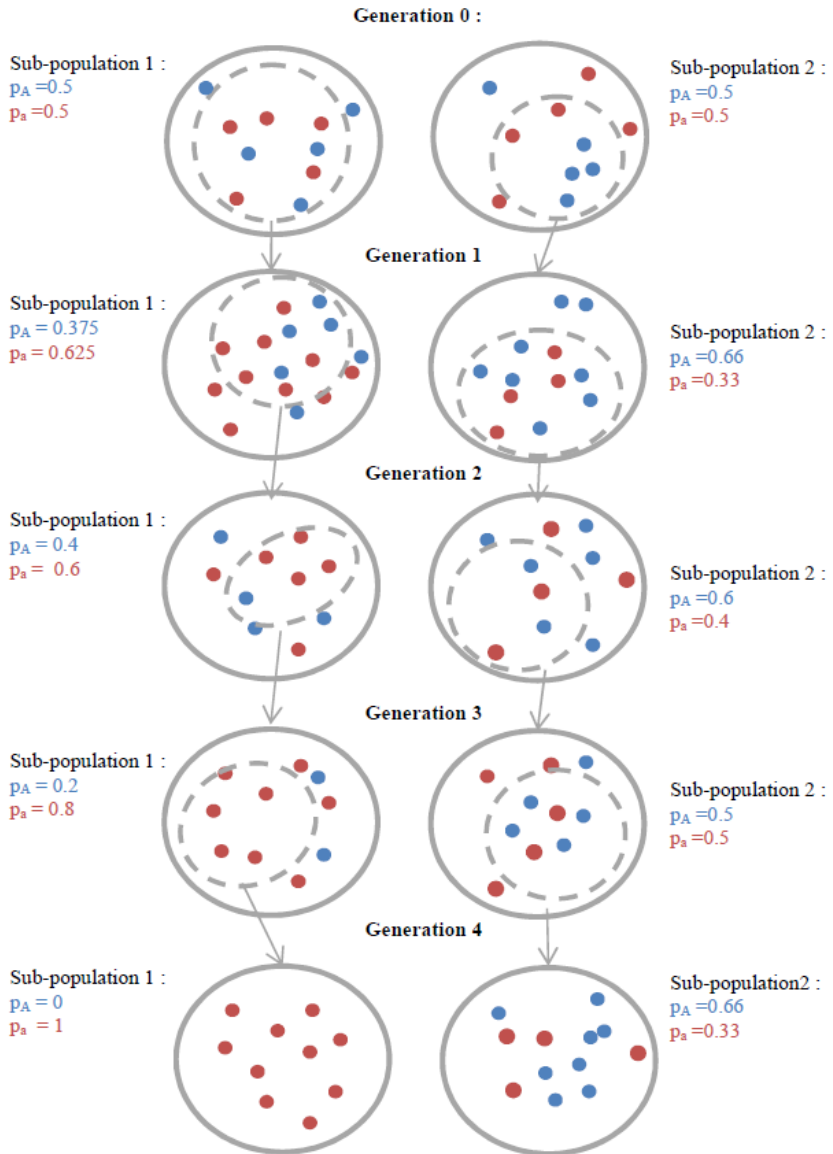


Figure 15: Random genetic drift in two sub-populations. The frequency of the alleles A and a, p_A and p_a respectively, may change through generations in each population independently from each other, resulting in the differentiation of the sub-population. At generation 4, the random genetic drift results, in the disappearance of the A allele in the sub-population 1.

The Wahlund effect refers to the reduction in heterozygotes caused by population structure (Wahlund, 1928). The Figure 16 illustrates how the occurrence of sub-populations in Hardy-Weinberg equilibrium results in a deficit of heterozygosity (diversity) in the all population. Moreover, the deficit increases proportionally with the differences in allelic frequencies among the sub-population.

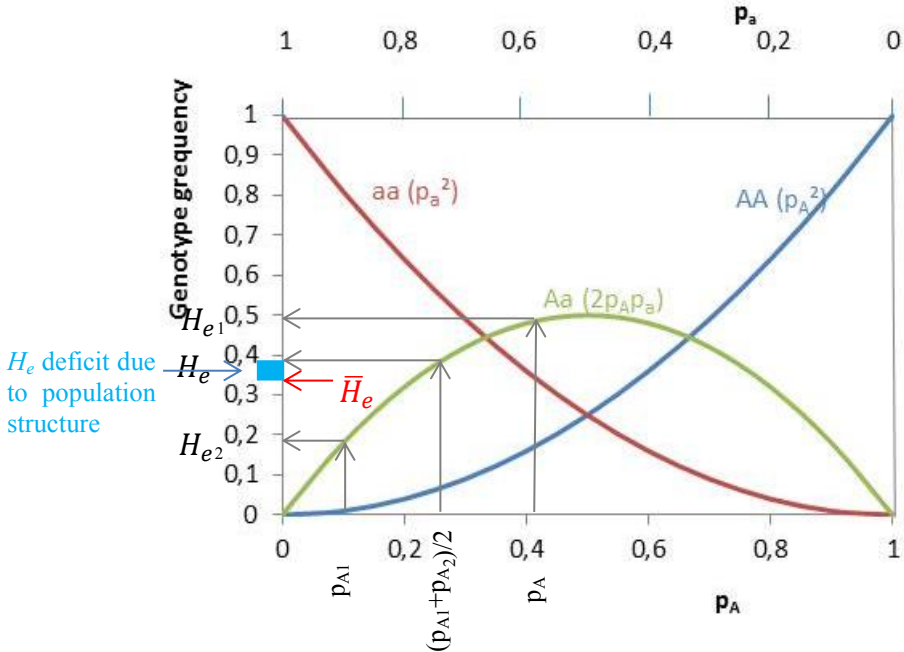


Figure 16: Deficit in heterozygote due population structure in a diploid population. Considering a population divided two sub-populations of the same size, with p_{A1} and p_{A2} the A allele frequencies in the first and second sub-populations respectively and p_{a1} and p_{a2} the a allele frequencies in the first and second sub-populations respectively, the A and a allele frequencies of the population are $(p_{A1}+p_{A2})/2$ and $(p_{a1}+p_{a2})/2$ respectively. In the case of a population without structure, the expected heterozygosity of the population is $H_e = (p_{A1}+p_{A2})(p_{a1}+p_{a2})$. In the case of several sub-populations with different allele frequencies, the expected heterozygosities in the two sub-populations are $H_{e1} = 2p_{A1}p_{a1}$ and $H_{e2} = 2p_{A2}p_{a2}$, and the actual heterozygote frequency observed in the population is $\bar{H}_e = (H_{e1} + H_{e2})/2$. Due to the shape of the $2p_A p_a$ curve, H_e is higher than $\bar{H}_e$. This difference increases with the differences in allele

frequencies of the two sub-populations.

Autogamy has an extreme effect on population structure where each individual is isolated. Mating systems favoring autogamy result in an increase in homozygote frequency compared to the Hardy-Weinberg expectation. Indeed, if a heterozygote individual ($p_A = 0.5$ and $p_a = 0.5$) is selfed, half of the F1 generation stemming from the selfing is heterozygote ($2 p_A p_a = 2 \cdot 0.5 \cdot 0.5 = 0.5$). In F2, half of the individuals stemming from the heterozygotes F1 are homozygotes and all the individuals stemming from the F1 homozygotes are homozygote. Therefore at the F2, the homozygote frequency is 0.75 while it would 0.5 under allogamy. Heterozygote frequency continues to decrease with the generation according to the formula $(\frac{1}{2})^{n-1}$. with n the number of the F generation.

The wright's F-statistics describes the distribution of the genetic diversity (heterozygosity) in a population and its sub-populations (Wright, 1965). In particular, the fixation index (F_{st}), computed as the relative deficit of heterozygote in the population due to the occurrence of sub-populations, provides a measure of the genetic differentiation among sub-populations or in other terms, to the proportion of the total genetic variation that separates the sub-populations, or to the proportion of genetic variance explained by the sub-population factor.

$$F_{st} = \frac{(H_e - \bar{H}_e)}{H_e}$$

H_e is the heterozygote frequencies of the population at Hardy-Weinberg equilibrium

$\bar{H}_e$ is the mean of the heterozygote frequencies in each sub-populations at Hardy-Weinberg equilibrium.

Fst ranges from 0 to 1 and increases with the differentiation of the allele frequencies of the different sub-populations.

As previously described, Fst may be used to evaluate the differentiation among sub-populations. The pairwise comparison of multiple sub-populations may be stored in a distance matrix. The term individual matrix distance is used in the extreme case where each sub-population corresponds to one individual only. Fst and other genetic distances are used to generate distance matrix.

Fst between groups of individuals may be evaluated as described above or using the analysis of the molecular variance (AMOVA). Alike the analysis of variance (ANOVA), the AMOVA dissects the variance components of a system, here the population structuration. Contrary to the ANOVA based on continuous data, the AMOVA computes the variance component by permuting dissimilarity matrix based on discrete parameters such as individual Fst distance matrix based on molecular markers. The AMOVA tables are read like an ANOVA tables describing the effect of the

sub-population on the genetic diversity. A p-value is assigned to the model and the F_{st} can be interpreted as R^2 . Different levels of structuration may be evaluated (sub-populations, and groups within the sub-populations). Here, we present the simple case of one level of population structuration. F_{st} is computed as:

$$F_{st} = \frac{(V_{Ap} - V_{wp})}{V_{wp}}$$

V_{Ap} variance among sub-populations

V_{wp} variance within sub-populations

Different statistical tools are used to identify and visualize the genetic relations among individuals or groups of individuals. The most popular are the multivariate, the clustering and the assignation analysis.

The multivariate analysis is a set of statistical methods designed to synthesize the variability of complex datasets constituted of elements described by numerous variables (Figure 17). The principle is to identify new variables (principal components, PC) that maximize the variance between the elements. The successive PCs explain successively a lesser part of the total variation. These PCs are linear combinations of the original variables and are characterized by their eigenvalue, a parameter that describes the variance of the original dataset explained by the PC.

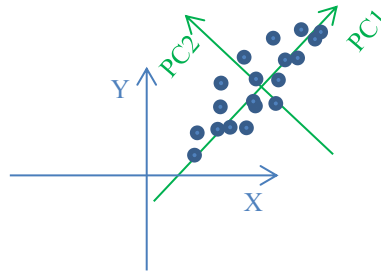


Figure 17: Multivariate analysis concept. The different elements are described by the variables X and Y. The first principal component (PC1) maximizes the variability of the elements. The second principal component (PC2) is independent (orthogonal) to the PC1 and maximizes the remaining variability. In a system including more variable, all PCs are independent from each other and the successive PCs explain successively a lesser part of the variation.

The most commonly used multivariate method is the principal component analysis (PCA), based on continuous variables. The principal coordinate analysis (PCoA) is similar to the PCA but explore dissimilarity matrix based on discrete variables such as Fst distances matrix based on molecular markers. Multivariate analysis does not involve any specific biological assumption. Finally, the distance separating the elements on the first PCs summarizes the distance based on several markers. Most often, the first two or three PCs reveal most of the genetic structure of a population.

The clustering analysis consists in the construction of dendrogram, i.e. branching diagrams showing the relationship among the different elements. Again the analysis is based on a distance matrix between individuals. The

sum of the branches separating two taxonomic units represents the distances separating them. The dendrogram construction is an iterative method including the repetition of three steps. The first one is the assembly of two existing taxonomic units in a new one. The second is the evaluation of the distance separating the new taxonomic unit and the other existing taxonomic units. The third is the evaluation of the length of the branches separating the new taxonomic unit and the two assembled taxonomic units. Two algorithms are commonly used in population genetics. (i) The unweighted pair group algorithm (UPGMA) that “simply” assembles pairs of taxonomic units presenting the lowest distance (Sokal, 1958). The distance between the new taxonomic unit and the other is the mean distance of the assembled taxonomic units and the others taxonomic units minus the distance separating the two assembled units. The length of the branches separating the two assembled taxonomic units is half the distance separating them. (ii) The Neighbor-Joining algorithm (NJ) is based on the parsimony principle i.e. evolution follow the shortest path (Saitou & Nei, 1987). Therefore, this more complex algorithm builds a tree that minimizes the total branches length. The branches length represents the number of mutation event since the separation the taxonomic unit. The number of mutation since the separation of two different element maybe different, therefore, the branch length separating two elements to their common taxonomic unit may be different. (Hamon et al. 1999, Hartl et al 2007)

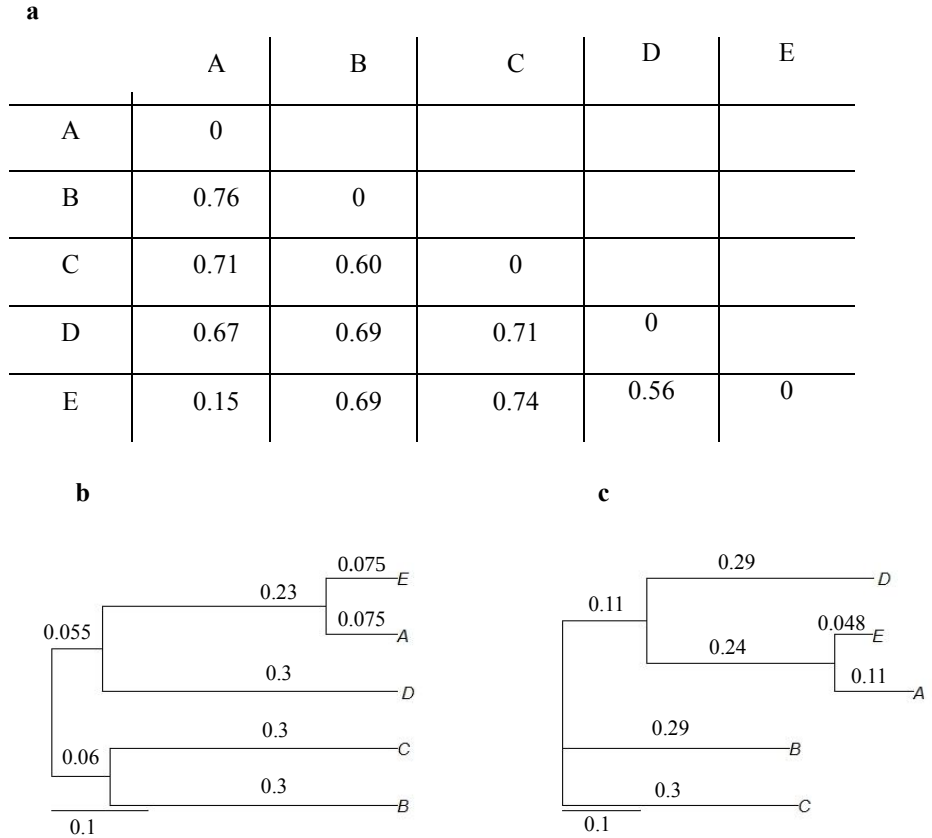


Figure 18: (a) Pairwise distance matrix between the individuals A, B, C, D and E. (b) UPGMA tree based on the pairwise distance matrix. (c) NJ tree based on the

pairwise distance matrix. On both trees, the distance between two individuals or clusters is the sum of the branches (horizontal segments) separating them. On both trees, E is the closest to A and C the closest to B. Nevertheless, the distance separating C and B (0.6) is higher than the distance separating E and A (0.15). E and A cluster together with D. On the UPGMA tree, the E-D and A-D distances are considered as equal but the NJ tree conserves the information that the distance separating E and D is smaller than the one separating A and D. The distances between the E-A-D and C-B group are smaller than the distance separating the individuals within these groups. This indicates a low differentiation among the two groups compare to differentiation between individuals within the groups.

The assignment method, implemented in the software Structure, aims at assigning the individuals to newly identified sub-populations, generally referred to as putative populations (Pritchard et al. 2000). The rationale is that the different sub-populations are in Hardy-Weinberg equilibrium. Therefore, for a defined level of population structure (K , the number of sub-populations), the assignment method uses an iterative sampling method - Monte Carlo Markov chain (MCMC) - to identify the K putative populations that maximize the Hardy-Weinberg equilibrium within the sub-populations. Due to the MCMC, for a predefined K , the different simulations lead to slightly different results. The quality of each simulation is described by two statistics: the likelihood and p-value. Therefore, the authors advise to run different simulation for each K and select the one presenting the best statistics. Besides, the assignment method does not define an “ideal” level of population structure (K) and the statistics describing the quality of the simulation increases with K (Figure 19). The Evanno method is a popular

method to identify the structure level (K) explaining the major component of the genetic structure (generally named best K) (Evanno et al. 2005). It is based on the rate of change in the $-\log(\text{likelihood})$ (L) compared to the variance among successive K: The best K is the one that provide the best $\Delta K = \text{mean}(|L''(K)|)/\text{sd}(L(K))$, with $L'' = |L'(K+1) - L'(K)|$ and $L'(K) = L(K) - L(K-1)$. This method is sensitive to the type of markers and number of loci used. Therefore I advise to correlate the different K results with available phenotypic and pedigree data. The Structure software included different models. Among the most popular models, one considers that each individual belongs to one putative population only and another allows some admixture in the individuals. In the latter case, a portion of the genome of each individual is assigned to the different putative populations. The population structure is more important when individuals are mainly assigned to one putative population and when the differentiation (measured by F_{st}) between the putative population is high.

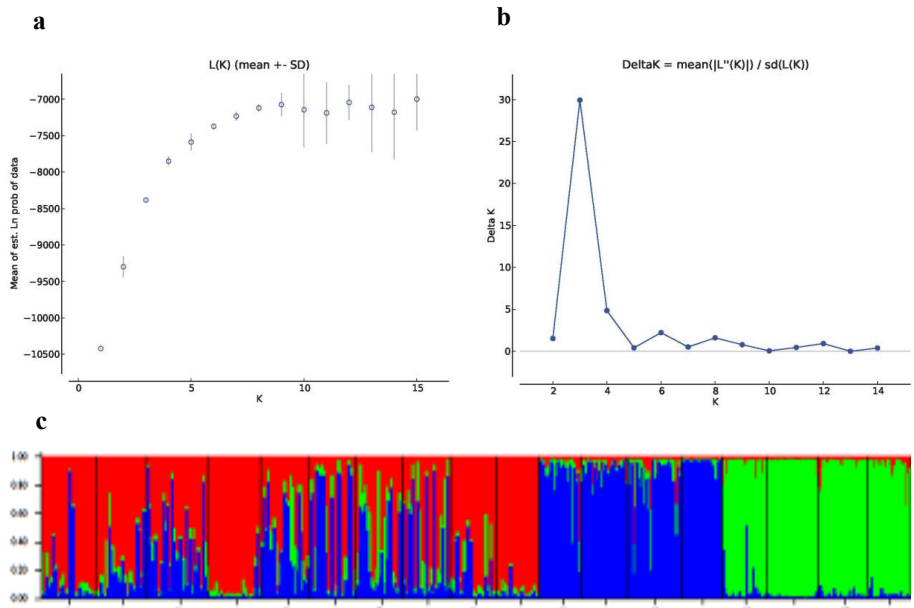


Figure 19: Graphical output of structure. (a) Typical graph of the mean value of $\log(\text{likelihood})$ ($L(K)$) according each structuration level (K). $L(K)$ gradually increases with K and reaches a maximum value. (b) ΔK for each structuration level (K). The peak at $K=3$ reveal the “best K ” according to the Evanno method. (c) Typical graphical output of Structure for $K = 3$. Each vertical line corresponds to one individual and the color corresponds to one of the three putative populations. When one color is predominantly represented in one individual, this individual is mainly assigned to this putative population. The individual presenting different colors have a mixed ancestry.

3.1.3 Linkage disequilibrium

The linkage disequilibrium (LD) is the non-random association of alleles at different loci (Figure 20). LD between loci is caused by their shared history of mutation and recombination (Flint-Garcia, Thornsberry, and Buckler 2003). Therefore, LD of a population mainly depends of the physical distance between the two loci: the LD (association) between two loci increases with the physical proximity. The LD extent describes the tendency of the LD to decline according to the distance separating the loci.

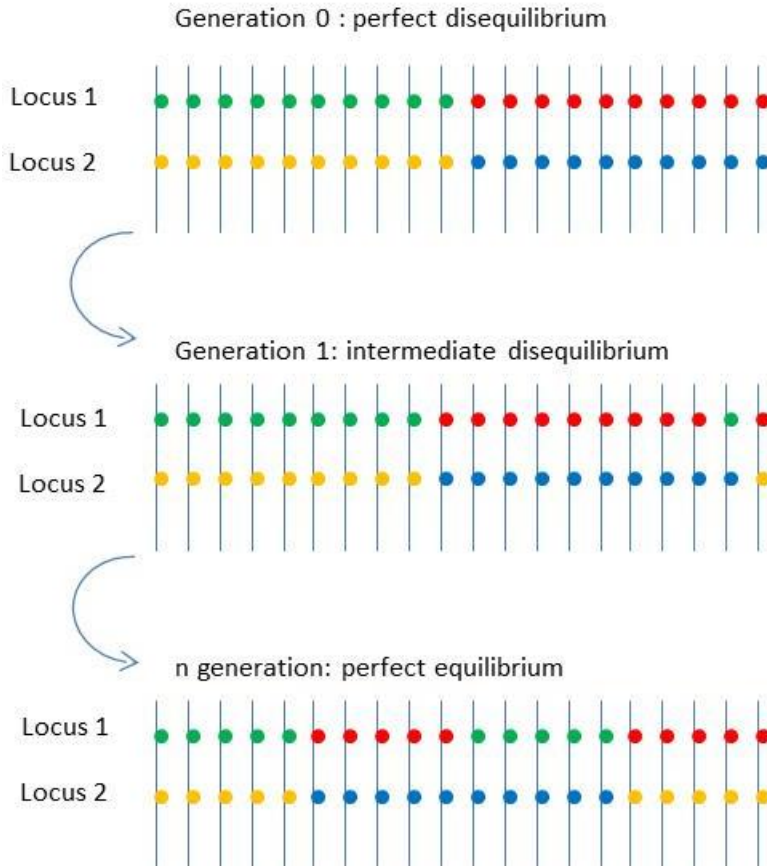


Figure 20: Theoretical representation of linkage disequilibrium (LD). Each line represents a haplotype with two biallelic loci. The first locus presents a green or a red allele, the second a yellow or blue allele. The generation 0 is constituted of two haplotypes (green-yellow and red-blue). The green allele is always associated to the yellow allele and the red allele is always associated to the blue allele. Therefore, the two loci are strongly associated: they are in perfect linkage disequilibrium. When this generation gives rise

to the generation 1, crossovers generate some new haplotypes (green-blue and red-yellow). Therefore four haplotypes are present in the generation 1. The green allele is preferentially associated to the yellow allele and the red allele is preferentially associated to the blue allele. Therefore the association between the two loci holds true but is weaker than in the generation 0: the two loci are in intermediate disequilibrium. The linkage disequilibrium (*i.e.* the association) between the two loci decreases with the generations. The number of recombinants would be higher if the distance between the two loci was higher. After n generations, the green and red alleles are equally associated to the yellow and the blue allele. There is no association between the two loci: they are in perfect linkage equilibrium, the two loci are considered as independent.

Many processes affect the LD. Originally, the LD arises from the appearance of a new allele in a haplotype, so that it is only associated to the other alleles neighbor present in this haplotype. Crossovers break down this association, therefore the most evident factor affecting LD are the mutation and recombination. The LD between pairs of markers in close proximity varies according to the age of the different alleles (number of cross-overs since the appearance of the new allele) and the average LD varies across the genome because some regions are more prone to recombination than others. Besides, average LD is a property of a population. Indeed, most of the processes observed in population genetics are also reflected in the LD pattern. Here are the main population factors affecting LD. Firstly, the mating system. Indeed, allogamous species, compared to autogamous species, tend to combine alleles originating from different haplotypes, resulting in a lower LD (Figure 21). Secondly, population structure corresponds to the occurrence of

sub-populations presenting different allele frequencies, so that in each sub-population, the alleles of different loci are preferentially associated to the alleles predominantly represented in the sub-population, generating higher extent of the LD of the population (Figure 22). Similarly, migration and admixture (population fusion) tend to increase the LD. Thirdly, a reduction in population size (bottleneck effect) randomly eliminates some haplotypes present in a population, so that it increases the association between loci. Fourthly, the selection increases the frequency of the alleles present in the haplotype carrying the selected allele, resulting in an increase in the LD extent in the genomic region surrounding the selected loci.

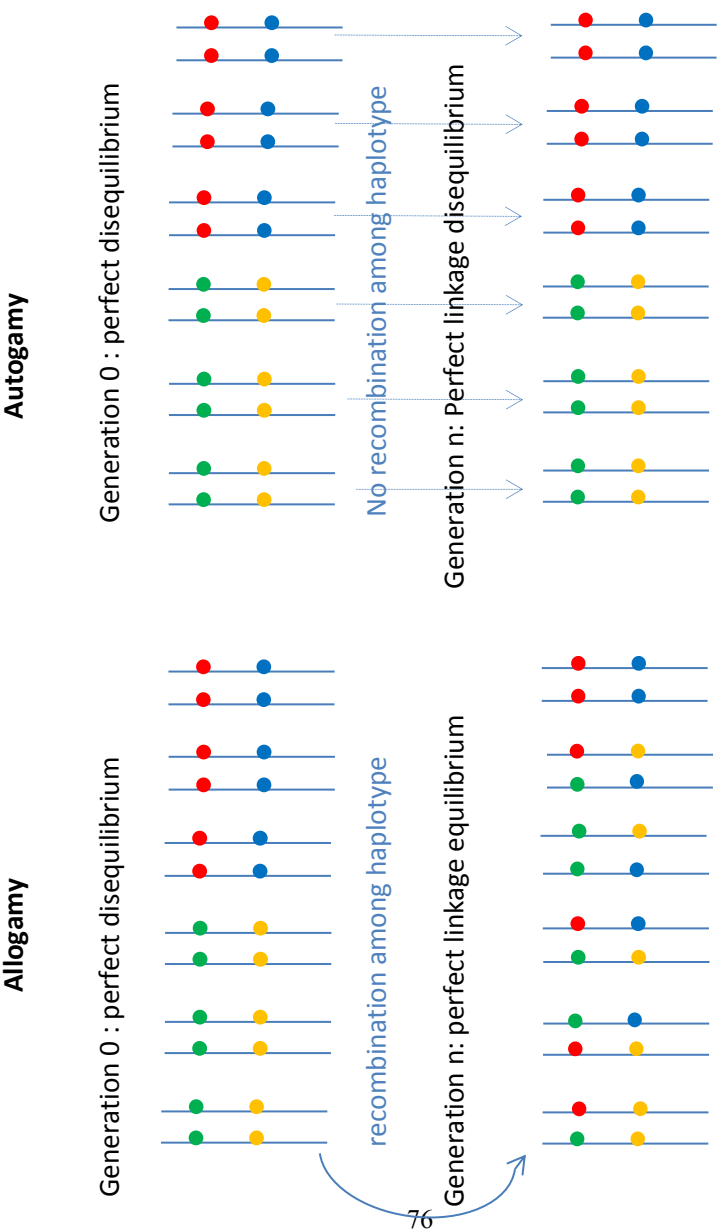
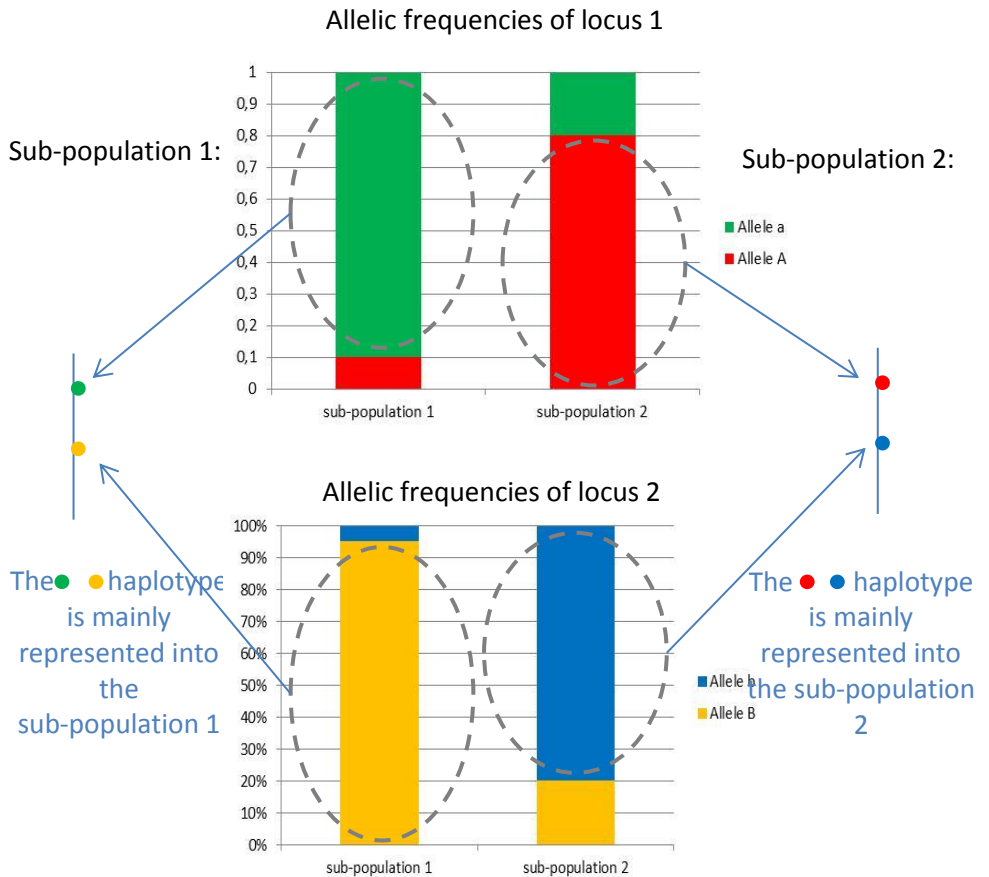


Figure 21: Effect of mating system on Linkage disequilibrium in diploid species. Each line represents a haplotype with two biallelic loci. The first locus presents a green or a red allele, the second a yellow or blue allele. The generation 0 is constituted of two haplotypes (green-yellow and red-blue) and homozygous individuals. The linkage disequilibrium is perfect. In allogamous mating systems (left), individuals breed with each other, the different haplotypes may be combined to generate new haplotypes in the following generation as described in the Figure 20. In autogamous mating system (right) The linkage disequilibrium is in the following generation remain perfect since each individual breeds with itself, so that the haplotypes transmitted to the next generation are a combination of the haplotype present in each individual. This limits the possibilities of generation of new haplotypes. Moreover, as previously described, selfing results in high level of homozygosity, so that most individuals contain two copies of the same haplotype, which limits the possibilities of haplotype combination.



Over representation of the a-B and A-b haplotypes compared to the haplotypes a-b and A-B



Linkage disequilibrium in the population constituted of the sub-population 1 and 2

Figure 22: Effect of population structure on Linkage disequilibrium in a diploid species. The histogram present the allelic frequencies of two loci in two sub-population. The first locus presents a green or a red allele, the second a yellow or blue allele. The green and yellow alleles are highly represented in the first sub-population, so that they are highly associated. Similarly, red and blue alleles are highly represented in the second sub-population, so that they are also highly associated. Finally, when the all population is taken into account, there is an over representation of the green-yellow and red-blue haplotypes compared to the expectation of no association of alleles into the population.

The LD measures aim at describing the deviation of the observed haplotype frequencies from their expectation assuming independence of the alleles from the different loci. Considering two biallelic loci with the alleles A and a at locus one, and B and b at locus two, with allele frequencies p_A, p_a, p_B, p_b and the haplotype frequencies $p_{ab}, p_{Ab}, p_{aB}, p_{AB}$ (Table 2). The basic component of all LD measures, the parameter D , is the difference between the observed and the expected haplotype frequencies under the independence of loci:

$$D = p_{ab} - p_a * p_b = p_{AB} - p_A * p_B = -p_{aB} + p_a * p_B = -p_{Ab} - p_A * p_b = p_{ab} * p_{AB} - p_{aB} * p_{Ba}$$

		Locus 1	
alleles		$A (p_A)$	$a (p_a)$
Locus 2	$B (p_B)$	$AB (p_{AB} = p_A p_B)$	$aB (p_{aB} = p_a p_B)$
	$b (p_b)$	$Ab (p_{Ab} = p_A p_b)$	$ab (p_{ab} = p_a p_b)$

Table 2 : Contingency table showing allele and haplotype frequencies of two bi-allelic loci under the assumption of independence of both loci. The locus 1 has two alleles A and a with the frequencies p_A and p_a respectively, the locus 2 the alleles B and b with the frequencies p_B and p_b respectively. The resulting haplotypes are AB, ab, aB and ab with the frequencies p_{AB} , p_{ab} , p_{aB} and p_{ab} respectively. Under the assumption of independence of both loci, the haplotypes frequencies are the product of frequencies of the alleles that constitute the haplotype.

The D parameter presents two main disadvantages. Firstly, the sign of D depend of the ordering of the alleles. Secondly, D depends of the allelic frequencies so that it is difficult to compare the D value of two pairs of loci presenting different allelic frequencies. Therefore different measurements were created to scale D . Among them r^2 is the most commonly used:

$$r^2 = D^2 / (p_a * p_A * p_b * p_B)$$

The r^2 corresponds to the square of the correlation coefficient between the two loci, so that it ranges from 0 (perfect linkage equilibrium) to 1 (perfect linkage

disequilibrium). Fisher test and permutations analysis may be used to evaluate the statistical significance (P-value) associated to the r^2 value.

The LD extent is generally represented by plotting the r^2 value according to the distance separating all pairs of available loci (Figure 23). As the LD values of loci separated by a same distance can present a wide range of values, the overall LD pattern may be estimated by fitting a non-linear regression through the points (Figure 23). Another representation is the LD heat plot that presents the LD variation across a chromosome (Figure 24). In this representation, the chromosome fragment including the markers is represented by a diagonal line. The color of each point of the plot corresponds to the r^2 value of the pair of markers that are on the right and below the point.

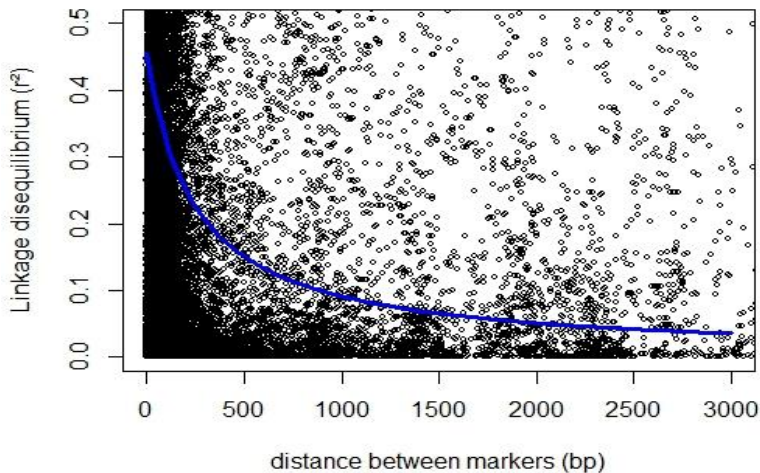


Figure 23 : Plot of the LD (r^2) according to the physical distance (base pair) between two loci. Blue line: non-linear regression based on pairwise LD values

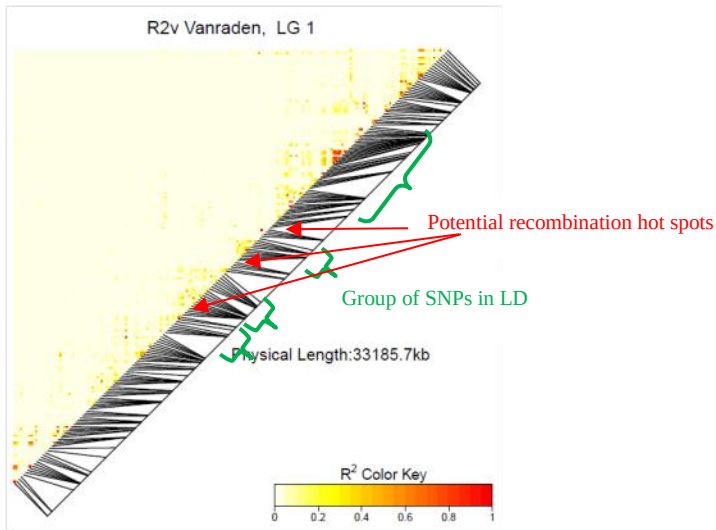


Figure 24: LD heat plot: the diagonal line represents a genetic map of genetic marker. Each colored point of the plot corresponds to the pair of loci (the loci below and on the right of the point) and its color corresponds to the r^2 value.

3.2 Linking genetic factors to phenotypes

3.2.1 Phenotype determinism

The phenotype is the result of genetic properties of an individual, the effect of the environmental conditions experimented by the individual (nutrition, climate, stress ...) and the interaction between these genetic and environmental factors. Indeed, all individuals do not react similarly to the same environmental conditions, for example, plants may be differentially affected by nitrogen availability. This is named genotype-environment interaction. Therefore, the phenotype may be modeled as the sum of the genetic, environmental effects and their interaction:

$$P = G + E + I_{GE}$$

P phenotype

G genetic effect

E environmental effect

I_{GE} interaction genotype-environment.

The genotype-environment interaction provides a high level of complexity so that it is not discussed in this document. According to these conditions, the phenotypic variance may be defined as follow:

$$Var(P) = Var(G) + Var(E)$$

Heritability is a statistics used to describe the fraction a phenotypic variability that can be attributed to the genetic variation. In particular, the proportion of phenotypic variance that can be attributed to the genetic variance, named broad sense heritability H^2 :

$$H^2 = \frac{Var(G)}{Var(P)}$$

The genetic effect may be dissected in the additive effect (A), the dominance effect (D) and the gene interaction (I_g).

$$G = A + D + I$$

$$Var(G) = Var(A) + Var(D) + Var(I_g)$$

Additive effect corresponds to the contribution of each allele independently of the others present in the genotype. Typically, when a biallelic locus presents only additive effects, the mean value of the heterozygote is equal to the mean values of the two different homozygotes. The dominance effect occurs when the contribution of each allele is affected by the other allele present on the same locus. Typically, the dominance effect corresponds to the deviation of the mean value of the homozygote compared to the value of the heterozygote. The gene interaction effect occurs when the contribution of an allele is affected by the allelic constitution at another locus. Therefore D and I are the properties of the genotype and the A effect is the property of the single allele. Considering that the distribution of the alleles in the gametes is at random, and since each parent

passes a single allele per locus in allogamous species, additive genetic variance is the component of the genetic variance responsible for similarity between individuals of two different generations.

Therefore, the proportion of phenotypic variance that may be transmitted from one generation to another under allogamy - the narrow sense heritability h^2 - is:

$$h^2 = \frac{Var(A)}{Var(P)}$$

3.2.2 QTL mapping and GWAS: an overview

QTL mapping and genome-wide association study (GWAS) exploit the information provided by a set of polymorphic markers spanning the complete genome. These methods require a mapping population, its genotyping with an high number of marker and its phenotyping for the trait of interest. The common rationale of both methods is that polymorphic marker loci in high LD (generally physically close to each other) with the genetic factors responsible for the variation of a given trait should be statistically associated with the phenotypic variation of the trait (Figure 25). In addition, this statistical association decreases with decreasing LD (i.e. with increasing distance between the marker locus and the causative factor). Through the systematic evaluation of this association for each marker locus, QTL mapping and GWAS delineate genomic regions most likely to contain factors responsible for the variation of the trait.

The LD extent within the collection determines the number of markers required and the accuracy of the localization of the causative loci. The number of genetic markers needed is inversely proportional to the extent of the LD. The size of the genomic region including the causative locus (QTL) is directly proportional to the extent of LD.

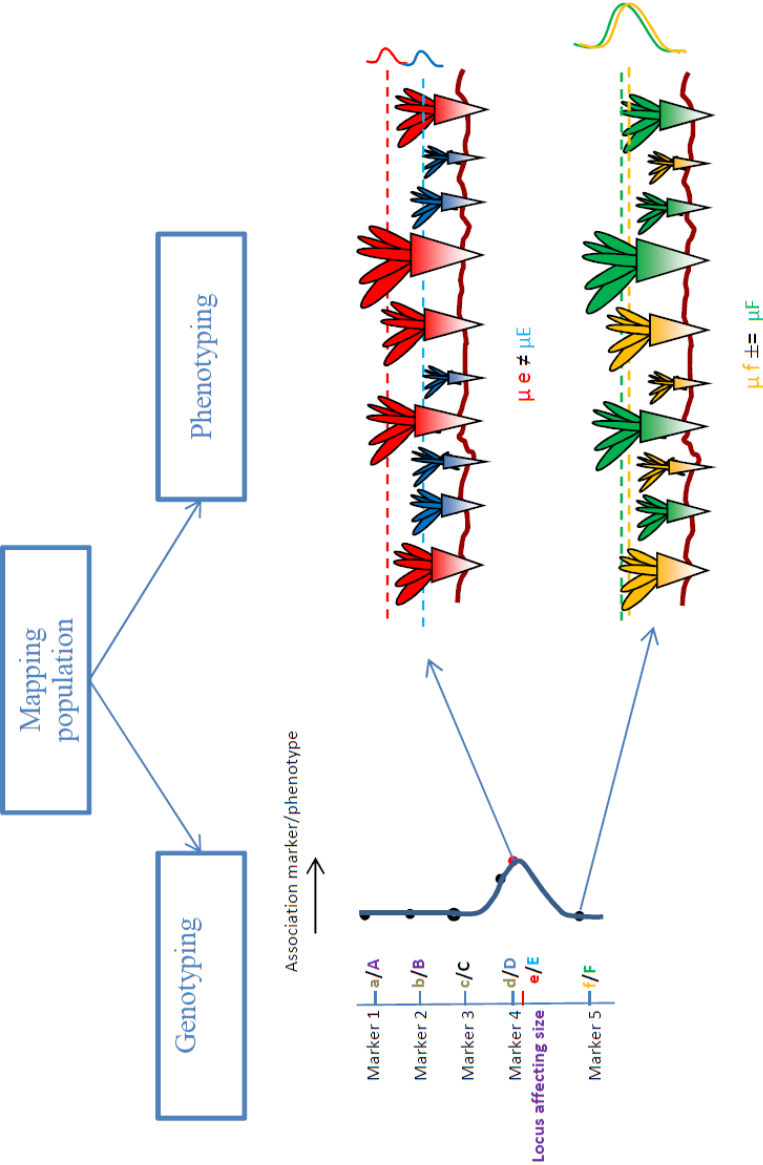


Figure 25: Schematic overview of GWAS and QTL mapping. The objective of both methods is to identify loci associated to one trait of interest. These methods require a mapping population, its genotyping with an high number of marker (left) and its phenotyping for the trait of interest (right). In this oversimplified example the trait of interest is the plant size and is controlled by one locus, the “locus affecting size” (purple), located between the markers 4 and 5. The association between the different markers and the phenotype are evaluated as follow: the mean of the individuals carrying one allele is compared to the mean of individuals carrying the other allele. In most cases, such as for marker 5, the mean of the individuals carrying the different alleles (f –yellow- or F –green-) are not significantly different from each other. Conversely, the individuals carrying the e allele (red) of the “locus affecting size” are significantly bigger than the individuals carrying the E allele. Since the marker 4 is physically close and in high LD with the locus affecting size, it also presents an association with the phenotype.

The difference between the two methods is the type of collection of individuals used. On the one hand, **QTL mapping** relies on experimental populations such as bi-parental F2 segregating population. These populations are characterized by a narrow genetic base, the absence of genetic structure and an LD that extends over long distances. The high extent of LD results in the identification of large QTLs (10 CM for a F2) using a relatively small number of genetic markers (100-1000). Due to the narrow genetic base of experimental populations, the causative genetic polymorphism existing in a species may be absent in a specific experimental population. The creation of an experimental population is also time consuming and may be difficult for some species.

On the other hand, **GWAS** exploits the genetic diversity present in a panel of unrelated individuals. The genetic diversity and the extent of LD are respectively higher and lower than in conventional experimental populations. The number of genetic markers needed is also considerably higher, from 5000 up to 10^9 . With populations showing a high LD extent, GWAS has a reduced resolution and allows the identification of large QTL-containing regions. However, provided a sufficient marker coverage of the genome, a populations presenting a very low LD extent, result in a more accurate the localization of the causative genetic factor is higher. Due to the large genetic base of diversity panels, the likelihood of a causative genetic polymorphism being present in a panel is generally higher than in an experimental population.

When prior knowledge concerning the genes related to the phenotype is available, GWAS can be speeded up by narrowing the evaluation of the association down to the set of polymorphic markers located within or close to the candidate genes. This approach is called a **candidate gene approach**.

3.2.3 GWAS

The association between each marker and the trait of interest is tested using an analysis of variance model. The basic statistical model is the simple way ANOVA:

$$Y = \mu + X\beta_0 + \varepsilon$$

Y : observed phenotypic value

μ : mean value

X : marker genotype

β_0 : effect of the marker

ε : error

The multiple origins of individuals gathered in a panel may result in the presence of population structure and family relationship among individuals. The population structure corresponds to the occurrence of sub-populations presenting different allelic frequencies for most of the loci (Figure 26). If some alleles affecting the trait of interest are also unevenly distributed in the sub-populations, all the loci that are unevenly distributed in the sub-groups will show some association with the trait of interest. This results in the identification of significant association that are attributable to the population structure instead of a physical proximity with the polymorphism at the origin of the trait variation, also called false positive. Within each sub-group, the family relationships among some individuals may result in the occurrence of identical chromosome fragments in these individuals. As for population structure, the phenotypic and genetic similarities between closely related individuals may lead to false association in the simple way ANOVA.

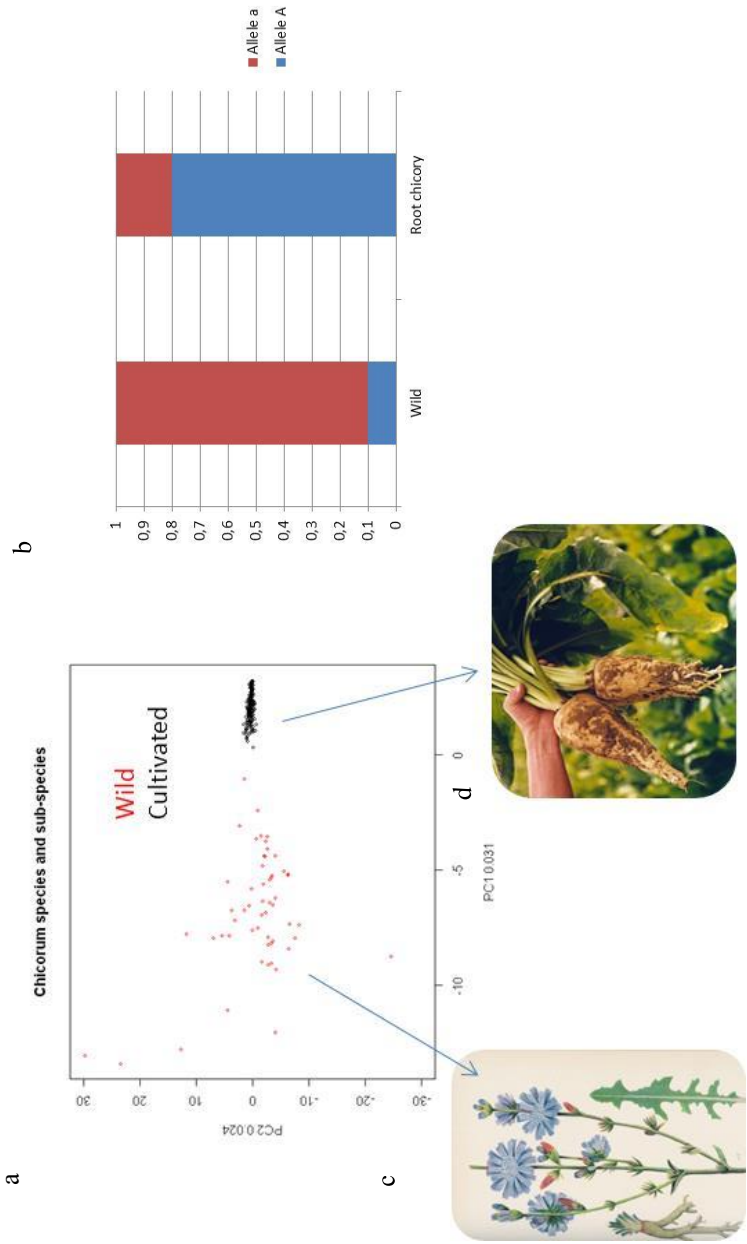


Figure 26 : Illustration of the effect of genetic structure on the association analysis. (a) PCA based on 15 SSRs markers presents the differentiation of a chicory population in two groups: the industrial chicories vs wild chicories. The separation into two groups mean that for many of the 15 SSRs markers and consequently, to many other polymorphic loci, the alleles present in the two groups are present in different frequencies and in most extreme cases, the alleles present into the two groups may be different from each other (b) Barplot illustrating the case of a typical biallelic marker with the alleles A and a in a structured population such it should be observed for a large number of markers in the chicory example. (c) Illustration of wild chicory. The wild chicory has a thin and disturbed root system and relatively narrow leaves with dented rib. (d) Picture of root chicory. The industrial chicory has a big taproot and large leaves. In the example, the allele A (blue on the barplot) could be statistically associated to most of the traits belonging exclusively to the wild chicories, such as the thin and disturbed root system and the relatively narrow leaves, when the a allele (red, on the barplot) could be associated to most traits characterizing the root chicory. In case of existence of population structure, many loci would have different frequencies in the two groups and therefore would present association with many traits that are different into the two groups. Nevertheless, in most cases, this association would be due to the genetic structure instead of a physical proximity of a genetic factor at the origin of the trait variation.

Mixed linear models are used to take into account the effect of the population structure (Q) and familial relatedness (or kinship relationship, K). The marker (X) and Q are generally analyzed as fix effect while K as random effect.

$$Y = \mu + X\beta_0 + Q\beta_1 + Ku + \varepsilon$$

Y : observed phenotypic value

μ : mean value

X : marker genotype

β_0 : marker effect

Q : structure matrix, it generally consists in PCoA or Structure results

β_1 : additive genetic effects associated the population structure

K : Kinship matrix, representing the familial relatedness using the inbreeding coefficient

u : random additive genetic effects associated to familial relatedness

ε : error

More complex models have been developed to integrate other components such as the environment effect (site, treatment) and the genotype-environment interaction.

For each marker, mixed linear models generate a p-value that describes the significance of the association and a R^2 describing the proportion of phenotypic variance explained for the locus. Due to the high number of markers tested in a GWAS, many markers can, by chance, present a significant association with the trait. Such a multi-testing problem may be solved by adapting (decreasing) the significant threshold (α) with the Bonferroni correction, which consists in dividing the signification threshold by the number of observation.

Beside the statistical threshold, the confidence allocated to an association is increased by the occurrence of several significant associations in a small DNA interval. Indeed, the markers surrounding a genetic factor affecting the trait tend to be in LD with genetic factor. Conversely an isolated marker may be potentially a false positive and should be considered with care. The most common representation of GWAS results is the Manathan plot that presents the $-\log(p\text{-value})$ against the position of the marker on the genome (Figure 27).

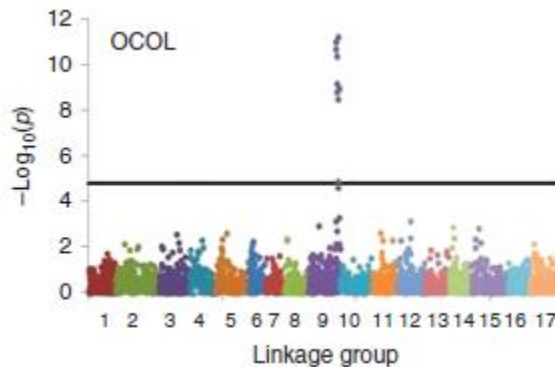


Figure 27: Manhattan plot presenting GWAS result, i.e. the level of significance ($-\log(p\text{-value})$) of the association between the different markers and a trait. The markers are ordered according to their position on the different chromosomes. Each chromosome is represented in a different color. Several markers present significant associations on the chromosome 9, revealing a QTL for the analyzed trait.

The quality of the family and population structure correction may be evaluated with a QQ plot that compares the distribution of p-values expected under the null hypothesis of no association and the observed p-values (Figure 28). A major deviation from the $X=Y$ line indicate that some factor, such as population

structure, environmental effect, affect the result of GWAS. A deviation of the extreme values of the QQ plot means that some significant SNPs were identified.

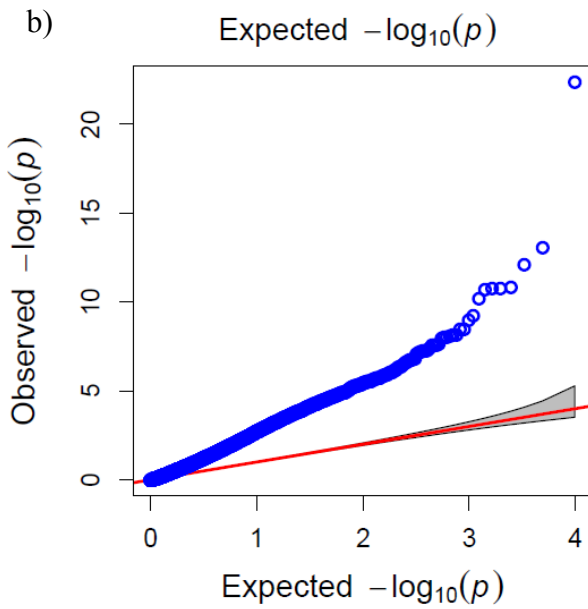
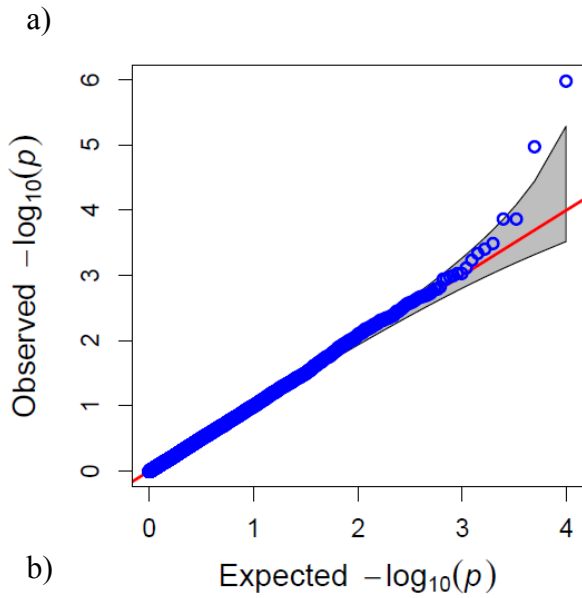


Figure 28: QQ plot of the GWAS results. In grey: confidence interval for the deviation. (a): QQplot with no major deviation from the $X=Y$ axes. This reveals that the model provides good correction for the population structure and family relatedness. The two points out of the confidence interval correspond to markers significantly associated to the trait of interest. (b) QQ plot of the result of a model that do not take into account the population structure. The deviation from the $X=Y$ axis indicate that many markers present highly significant values due to the population structure, family relatedness or another variable that should be integrated into the GWAS model.

The section concerning the quantitative genetic was inspired by Falconer & Mackay 1996, Lynch & Walsh 1998, Oraguzie et al. 2007, Mackay et al. 2007. Zhu et al. 2008.

3.3 References

- Acqua, H. G. (2012). Principle of plant genetic breeding. *John Wiley & Sons*.
- Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, 14(8), 2611-2620.
- Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131(2), 479-491.
- Falconer, D. S., Mackay, T. F., & Frankham, R. (1996). Introduction to quantitative genetics (4th edn). *Trends in Genetics*, 12(7), 280.
- Flint-Garcia, S. A., Thornsberry, J. M., & Buckler I.V. (2003). Structure of linkage disequilibrium in plants*. *Annual review of plant biology*, 54(1), 357-374.
- Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 49-50.
- Hartl, D. L., Clark, A. G., & Clark, A. G. (2007). Principles of population genetics. *Sunderland: Sinauer associates*.
- Lynch, M., & Walsh, B. (1998). Genetics and analysis of quantitative traits (Vol. 1). *Sunderland, MA: Sinauer*.
- Mackay, I., & Powell, W. (2007). Methods for linkage disequilibrium mapping in crops. *Trends in plant science*, 12(2), 57-63.
- Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12), 3321-3323.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population

structure using multilocus genotype data. *Genetics*, 155(2), 945-959.

Rikkerink, E. H., Gardiner, S. E., & De Silva, H. N. (2007). Association mapping in plants. *New York: Springer*.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4), 406-425.

Sokal, R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38, 1409-1438.

Thiagarajan, R. (2014). Principles of Animal Genetics and Population Genetics *Satish Serial Publishing House*.

Wahlund, S. (1928). Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas*, 11(1), 65-106

Weinberg, W. (1908). Über vererbungsgesetze beim menschen. *Molecular and General Genetics MGG*, 1(1), 440-460.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, 395-420.

Zhu, C., Gore, M., Buckler, E. S., & Yu, J. (2008). Status and prospects of association mapping in plants. *The plant genome*, 1(1), 5-20

4 Thesis outline

The general objective of this thesis is to unravel the genetic determinism of the synthesis of carbohydrates (including inulin) and the cold-induced inulin degradation kinetics of inulin in industrial chicory.

4.1 Specific objectives

The analysis of genetic determinism of a phenotypic trait relies on the premises that the genetic variability in the germplasm of interest contributes significantly to the phenotypic variability. The first specific objective of the thesis will be to characterize the genetic variability in the *Cichorium intybus* germplasm, with a particular focus on industrial chicory germplasm.

Before this thesis project, Nicolas Dauchot and Christine Notté investigated the association, in a small experimental panel, between, on one side, the carbohydrate (including inulin) quality at harvest and the susceptibility to cold-induced degradation and, on the other side, the polymorphisms within genes of the carbohydrates metabolisms, belonging to the GH 32 family. Among them, some polymorphisms in fructan hexohydrolases genes (FEH) appeared to be associated to a high proportion of the phenotypic variation of the cold-induced inulin degradation. No polymorphism could be associated to the carbohydrate quality at harvest. Besides, the genes associated to major effect QTLs in wheat and onion appeared to be closely or directly associated to genes

of the fructan metabolism. Based on these informations, we hypothesized the genetic determinism carbohydrate (including inulin) quality at harvest is determined by an array of minor effect QTLs, while that of cold induced degradation is determined by a combination of few major effect QTLs and many minor effect QTLs. Accordingly, the second and third specific objectives of the thesis will be, respectively, to confirm the presence of major effect QTL accounting for the cold-induced degradation and to identify minor effect QTLs associated carbohydrate quality at harvest and cold-induced degradation. Unlike prior studies in chicory, these hypotheses will be tested using a broad chicory germplasm, characterized in the frame of the first specific objective.

4.2 Strategy

To achieve these three specific objectives, we adopted a strategy with the following seven work packages (WP).

- I. Characterization of the genetic diversity available in chicory;**
- II. Creation of a diversity panel suitable for candidate-gene approach and GWAS;**
- III. Phenotyping of the diversity panel (field trials);**
- IV. High density genotyping of candidate genes.**
- V. Genome wide genotyping of the diversity panel;**
- VI. Candidate gene investigation for the susceptibility to cold induced inulin degradation;**
- VII. Genome-wide investigation of genetic factors associated to**

variation in inulin synthesis and cold induced-harvest degradation.

The following sections provide more details of the different working packages.

I. Characterization of the genetic diversity available in chicory;

The aim of this WP is to characterize the genetic diversity available in the *Cichorium intybus* L. germplasm in order to create a diversity panel suitable for the candidate gene approach and the genome-wide association study. Using 15 SSR markers, we investigated the genetic diversity of a broad collection spanning the different chicory types (wild, leaf cultivars, Witloof and root cultivars). Given the objectives of the thesis, the collection was deliberately enriched in the industrial root type: a total of 1297 individuals from the 15 cultivars at the origin of the breeding of the industrial root chicory were genotyped during the project.

The exploration of the genetic diversity was carried out in three sampling steps, corresponding to three growing seasons (2011 to 2013). The first sampling included an even representation of the different cultivars and root types and led to the definition of groups of genetically related entities. The second and third samplings were carried out on industrial cultivars and deepened the analysis of, respectively, the largest group (9 French cultivars) and the remaining smaller groups (6 cultivars from others origins) (Figure 29).

II. Creation of a diversity panel suitable for candidate-gene

approach and GWAS;

The first sampling in WP1 revealed the existence of a large genetic diversity both in the chicory germplasm and the industrial chicories, making this material suitable for the second and third objectives of the thesis. Based on these results, we assembled a diversity panel used for both the candidate gene and genome-wide approaches. This panel was restricted 15 cultivars at the origin of the current breeding program of industrial chicories. Indeed, wild individuals and leaf chicories present a very different root shape (thin and branched roots instead of a big taproot). Therefore, their QTLs associated to carbohydrate synthesis and inulin degradation are likely to be different from that of industrial root type.

During each of the three growing seasons, approximately 125 individuals were selected and integrated into the diversity panel. The selection was based on the SSR genotype in order to maximize the genetic diversity of the diversity panel. After the third season, the panel included a total of 373 individuals.

III. The phenotyping of the diversity panel;

In WP3, we performed the field trials in order to observe the phenotypic values (e.g. carbohydrate quality and inulin degradation) of each individual of the diversity panel. Since the carbohydrate analysis is a destructive analysis and because the individuals of the panel cannot be cloned, we implemented a progeny-test strategy that consisted in selfing the individuals of the panel and phenotyping their progeny during the next season. In the thesis, the term “line” refers to the progeny of an individual of the panel.

For each line, the carbohydrate analysis was performed at harvest on a part of the progeny and after a storage period on the rest of the progeny. This procedure simulated the cold temperatures occurring in the field at the end of the growing season.

IV. The candidate genes typing in the diversity panel;

The candidate gene analysis (WP VI) was carried out on subsets of the panel that were genotyped for different genes of the GH32 family. Before the start of this thesis several indels and SSRS markers had been identified in different genes of the GH32 family by Nicolas Dauchot. In the frame of WP4, we identified new indels as well as copy number variant (CNV) polymorphisms in the same gene family. The SNPs and indels were genotyped for each individual of the panel selected during the first growing season. Due to the prohibitive cost, the CNV typing was realized on a subset of 36 individuals.

V. The high density and genome wide genotyping of the diversity panel;

In order to provide the genotypic information required for GWAS (WP7), each individual of the panel has been genotyped using genotyping-by-sequencing (GBS). This method aims to sequence an identical fraction of the genome by sequencing the flanking regions of a restriction enzyme site in each individual. It leads to the identification of a large number of polymorphisms within these sequences.

VI. The validation of the association between the susceptibility to inulin degradation and the candidate gene;

Candidate gene approach tested the association between, on one side, the polymorphisms identified in the working packages IV and, on the other side, different traits describing carbohydrate quality at harvest and the susceptibility to cold-induced inulin degradation. The candidate-gene approach was implemented on a subset of the panel corresponding to the individuals selected during the first growing season.

VII. The identification of genetic factors associated to variation in fructan synthesis and susceptibility to inulin degradation

The genome-wide investigations of polymorphisms associated to carbohydrate quality and susceptibility to inulin degradation was realized by genome-wide association mapping. The GWAS was preferred to QTL because of its higher accuracy and was possible given the high number of markers generated by the GBS.

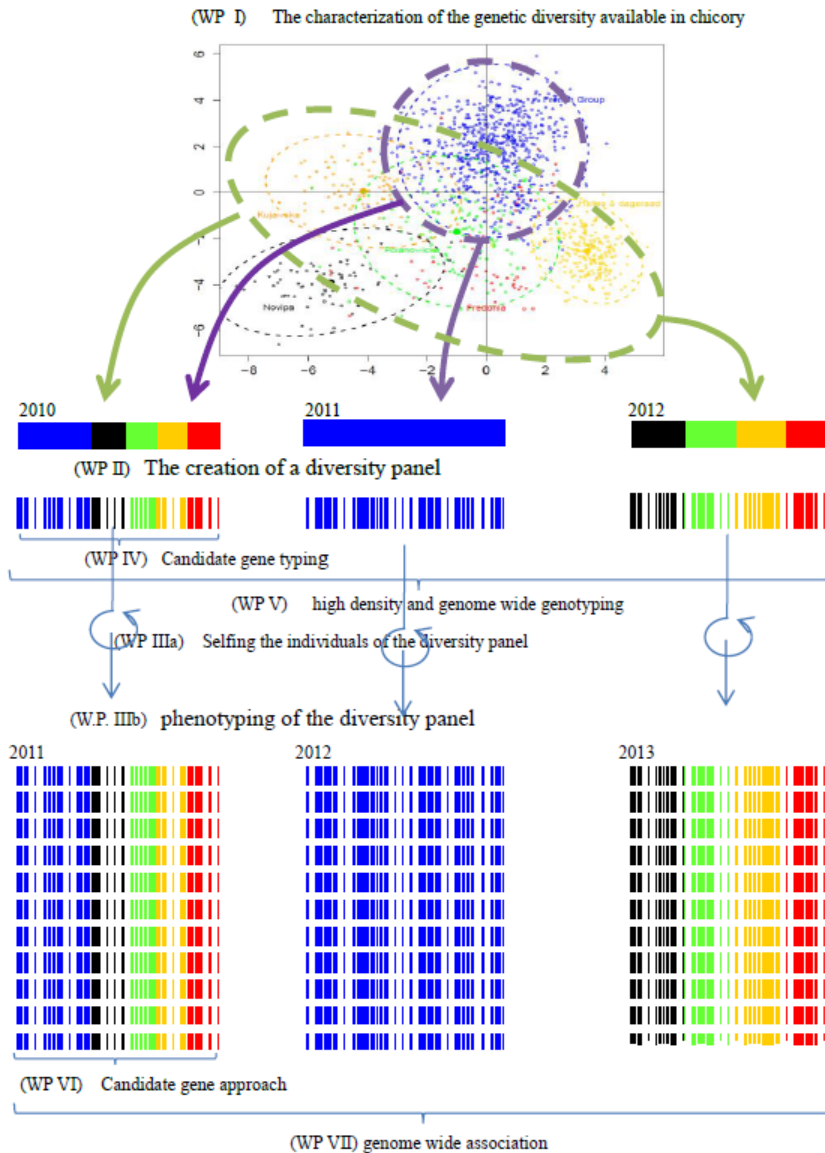


Figure 29: Schematic view of different worki packages (WP) and the repartition of the individuals across work packages and time. WP I: The characterization of the genetic diversity available in chicory. PCA of the collection of industrial chicories. The blue points correspond to the individuals of the 9 French cultivars. The points in black, green orange and red correspond to individuals belonging to other cultivars. Each color corresponds to differentiated cultivar groups. The genotyping of the individuals was dispatched on three growing seasons as follow: 2010: individuals from all cultivar groups, 2011 individuals belonging to the 9 French cultivars , 2012 individuals belonging to the 6 other cultivars only. WP II: The diversity panel was selected among the collection of individuals each vertical line corresponds to one individual and the color to its cultivars group. WP III: individuals of the diversity panel were (a) selfed and (b) their progenies grown and phenotyped for their carbohydrate content during the newt growing season. WP IV: A subset of the panel corresponding to the individual selected the first growing season (2010) was genotyped for candidate genes polymorphisms. W.P.V: All individuals of the panel were genotyped by the GBS technique. WP VI: Candidate gene approach based on the subset of the panel corresponding to the individual selected the first growing season 2010) WP VII: Genome wide association based on the entire panel.

4.3 Thesis organization

The thesis is organized in four research chapters (Figure 30):

First research chapter: Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the gene pool of industrial chicory

This chapter aims at characterizing the genetic resources available for the breeding of industrial chicory (WP I). We clarified the relationships of industrial chicory within the *C. intybus* germplasm and investigated the genetic diversity and structure of the original germplasm used to breed the modern cultivars of industrial chicory.

Second research chapter: Mutations in chicory FEH genes are statistically associated with enhanced resistance to cold-induced inulin depolymerization

This chapter describes a candidate-gene approach among inulin and sucrose synthesis and degradation genes, all members of the family 32 of glycoside hydrolases (GH32). This was achieved in two steps. Firstly, polymorphisms within the GH32 family were identified by eco-tilling on two groups of chicory breeding lines exhibiting contrasted response to cold-induced inulin degradation (WP IV). Secondly the association between polymorphisms within the FEH genes was then confirmed on a panel of 116 lines (WP II,

III and VI).

Nicolas Dauchot realized most of the writing and produced most of the result presented in this article. Along with C. Notté and O. Maudoux, I created the panel of 116 individuals used in the document.

Third research chapter: Loss of function of 1-FEH IIb has more impact on cold-induced inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa

The third chapter comes in the direct continuity of the chapter 2 with the identification of the actual 1-FEH IIa polymorphism the origin the loss of sensitivity to the cold-induced inulin degradation, and therefore the identification of the underlying mechanism (WP IV and VI).

Again, Nicolas Dauchot realized most of the writing and produced most of the result presented in this article. I realize most of the statistical analysis concerning the phenotype and molecular markers.

Fourth research chapter: GWAS in chicory identified SNPs associated to the carbohydrate and inulin quality.

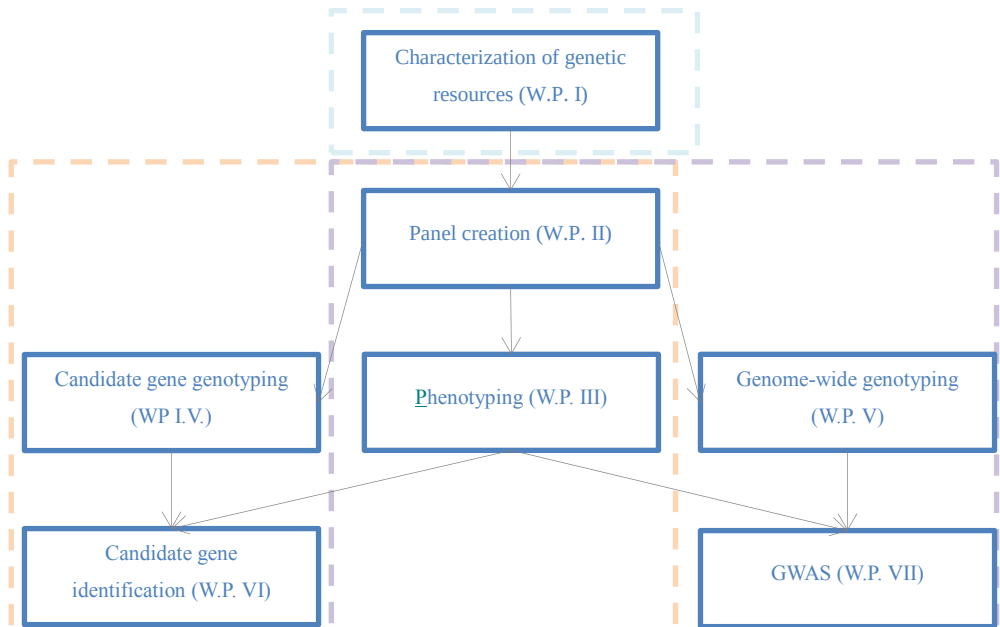
This chapter describes the genome-wide association mapping work on a panel of 366 lines genotyped by GBS and phenotyped for (i) the carbohydrates quality at harvest and (ii) their susceptibility to cold-induced inulin degradation. The genotyping generated 122 000 SNPs, part of which

were used to evaluate the linkage disequilibrium present in the industrial chicory germplasm and identify some SNPs associated to variation in carbohydrate synthesis and cold-induced inulin degradation (WP II, III, V and VI).

This chapter responds to the specific objectives II, III, V and VII.

Specific objective 1:

Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the gene pool of industrial chicory



Specific objective 2:

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest inulin polymerization

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa

Specific objective 3:

GWAS in chicory identified SNPs associated to the carbohydrate and inulin quality

Figure 31: Thesis outline and repartition of the specific objectives and working packages among the chapters.

This thesis integrates the project DGA INULONG, funded by the *Région Wallone*. Three partners are involved in the project: (i) Chicoline, a breeding company based in Warcoing (Belgium), (ii) the URBV lab (Plant Cellular and Molecular Biology) from the Université de Namur and (iii) the Earth and life institute-Agronomy (Elia) from the Université catholique de Louvain (UCL). In this project, I was in charge of the phenotyping, the quantitative genetic analysis and the statistical analysis. I also contributed to some genotyping work (GBS) with URBV. The URBV performed most of the molecular biology and Chicoline conducted the field experiments and phenotyping.

5 Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the gene pool of industrial chicory

This chapter is an adapted version of the research article entitled:

Raulier, P., Maudoux, O., Notté, C., Draye, X., & Bertin, P. (2015). Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the gene pool of industrial chicory. *Genetic Resources and Crop Evolution*, 1-17.

Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the
gene pool of industrial chicory

5.2 Abstract

The present study used 15 simple sequence repeat loci to characterize the genetic diversity of the germplasm that originated the current industrial chicory and to establish the relationships between and inside *Cichorium intybus* L. and *Cichorium Endivia* L. Initially we analyzed 19 cultivated *C. endivia* accessions, 27 wild and 155 cultivated *C. intybus* accessions distributed among three groups: 83 root chicories, 42 Witloof and 30 leaf chicories. The leaf chicories comprised cultivars corresponding to the Radicchio, Sugarloaf and Catalogne subgroups. The latter has not been previously included in any genetic diversity study. Subsequently, 1297 individuals from the 15 modern root chicory cultivars at the origin of the breeding of the current industrial root chicory cultivars were analyzed.

Although the accessions of *C. endivia* and *C. intybus* were clearly separated from each other, seven wild *C. intybus* individuals were genetically closer to *C. endivia* than to *C. intybus*, revealing complex genetic interrelationships between these species. The differentiation of *C. intybus* into three cultivar groups (Witloof, root chicory and leaf chicory) was confirmed. The leaf chicory individuals were divided into three genetic subgroups, corresponding to the Radicchio, Sugarloaf and Catalogne cultivars, thus attesting to the validity of the classification based on morphological factors. Clear differentiation was observed among the Belgian, Polish and Austrian modern industrial root cultivars, but not among the French industrial modern root cultivars. The high phenotypic and genetic variability of the modern industrial root cultivars indicates that this

Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the
gene pool of industrial chicory
germplasm constitutes a useful gene pool for cultivar improvement and selection.

5.3 Introduction

The *Cichorium* genus contains six species, of which four are only wild (*Cichorium botae* Deflers., *Cichorium spinosum* L., *Cichorium calvum* SCH.Bip ex Asch , and *Cichorium pumilum* Jacq), one is only cultivated (endive, *Cichorium endivia* L.), and one contains both cultivated and wild individuals (chicory, *Cichorium intybus* L.). Both cultivated species count numerous cultivars of leaf vegetables that are eaten raw or cooked. Some chicory cultivars are characterized by a large taproot and were originally used as a coffee substitute but are now mainly used for inulin production.

Recent studies (Kiers et al. 1999, 2000; Gemeinholzer and Bachmann 2005) have resolved phylogenetic relationships within *Cichorium* (Figure 32). *C. botae* is well differentiated from the other *Cichorium* species. The remaining species separate into two groups, of which one includes *C. intybus* and *C. spinosum*, and the other comprises *C. endivia*, *C. pumilum*, and *C. calvum*. Gemeinholzer and Bachmann (2005) could not genetically discriminate *C. intybus* from *C. spinosum* with three independent marker sets: *i.e.* nuclear ribosomal internal transcribed spacer (ITS), amplified fragment length polymorphisms (AFLPs), and simple sequence repeats (SSRs). Nevertheless, these authors observed that the morphological traits discriminating the two species are fixed and heritable, so that the morphological criteria for species delimitation are fulfilled. They concluded that *C. intybus* and *C. spinosum* are at an early stage of the speciation process. Apart from *C. botae* and *C. spinosum*, phenotypic delimitation of *Cichorium* species is difficult because it is mostly based on the life cycle, the

compatibility system, and highly variable quantitative features such as pappus scale characters (Kiers et al. 2000). *C. botae* presents cushion-like growth form, thus it is perfectly adapted to the warm, arid environment in its distribution area (Yemen and Saudi Arabia). *C. spinosum* can be identified by its spiny terminal branches. *C. intybus* has long pappus scales, is mainly self-incompatible and biennial. *C. endivia*, *C. pumilum*, and *C. calvum* are all annual and self-compatible. Species delimitation is complicated because cross pollination may occur between *C. intybus* and *C. spinosum*, and between *C. intybus* and *C. endivia* (Rick 1953; Vermeulen et al. 1994; Kiers 2000).

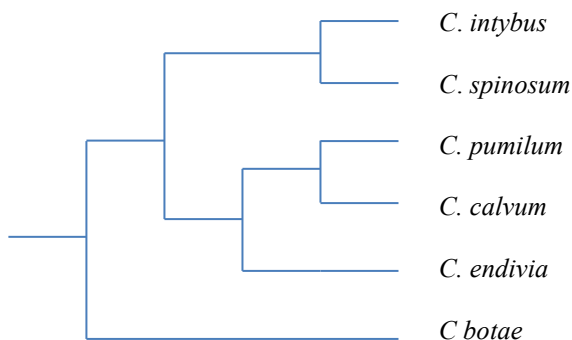


Figure 32: Subset of a single parsimony tree based on AFLP from Kiers 2000

No wild relative is known for *C. endivia*. Three cultivar groups are traditionally distinguished within this species (the Scarole, Frisée, and Endivia groups). However, this classification is not supported by AFLP analyses (Kiers 2000).

Both wild and cultivated forms are currently known for *C. intybus*. AFLP analyses show that they are genetically differentiated from each other (Van Cutsem et al. 2003; Sorensen et al. 2007). Several classifications of the different chicory cultivar groups have been described. Here, we present the classification of Kiers (2000), because it is supported by AFLP and ITS analyses (Kiers 2000; Van Stallen et al. 2001). This classification system divides the cultivated forms of *C. intybus* into three groups (leaf chicory, root chicory, and witloof), as outlined below (Figure 33).

- 1) The leaf chicory group consists of leafy vegetables that can be consumed fresh or cooked. The classification of Kiers divides the group into two subgroups: Sugarloaf and Radicchio. The sugarloaf subgroup is also referred to as *C. intybus* var. *porphyreum* and the Raddichio subgroup as Pain de sucre or as *C. intybus* var. *latifolium* (refer to Hammer et al. 2013 for more details). The Sugarloaf subgroup comprises several cultivars of green-leaved vegetables, producing a firm heart, and with a rather bitter taste. This subgroup is mainly cultivated in north-western Europe. The Radicchio subgroup originated in Italy and includes cultivars of various leaf colors (green, blond, and red). It is worth noting that the molecular relationship between the Sugarloaf and Radicchio subgroups has been poorly investigated because analyses of the relationship among them have included only three accessions: two Radicchios and one Sugarloaf (Kiers et al. 2000) or only one Sugarloaf accession (Kiaer et al. 2009).

A third subgroup of leaf chicory is described by different authors. It was designed as the Catalogne by Lucchin et al. (2008), as *C. intybus* var. *intybus* or *C. intybus* var. *sylvestre* by Hammer et al. (2013). Some authors (Van Cutsem et al. 1993) and seed breeders (e.g. Graines Grelin, Franchi Sementi, and LITO BM) also refer to this group as asparagus chicory or wild chicory, leading to a general confusion concerning this group. In this document, this group is referred to as the Catalogne subgroup. The Catalogne subgroup is characterized by thin, deep green, upright leaves, an indented or continuous lamina, a conspicuous midrib, and the absence of a firm heart. Therefore, the Catalogne subgroup is phenotypically distinct from the Radicchio and Sugarloaf subgroups. It is recognized as being of Italian origin. To our knowledge, accessions of the Catalogne subgroup have not been included in a previous study of genetic diversity.

- 2) The root chicory group, also known as industrial chicory or *C. intybus* var. *sativum*, is characterized by a large tap root, similar to that of sugar beet. It was originally grown as a coffee substitute, but is now mainly grown for inulin production. In the 1970s some cultivars were developed for forage in New Zealand (Rumball et al. 1986). Koch and Jung (1997) investigated the phylogenetic relationship of 17 cultivars using RAPDs and AFLPs. The DNA of four to eight individuals per cultivar were pooled. However, no clear grouping among the cultivars was identified, probably because of the sampling strategy and the suspected high level of intravarietal variability. Nevertheless, a high level of genetic diversity was detected, which provides potential for

future selection of this open-pollinated crop. In Italy the root of ancient varieties such as ‘Magdeburg’, ‘Brunswick’ and ‘Pont de Pierre’ are consumed as vegetable when harvested young and tender (Hammer et al. 2013).

- 3) The Witloof group, also referred as Belgian endive, Chicon or *C. intybus* var *foliosum*, produces witloof, a vegetable composed of tightly packed white leaves forced under artificial conditions. Witloof is thought to have originated around 1850, in the Jardin botanique de Bruxelles, Belgium, from the cultivar ‘Magdeburg’, an old root chicory cultivar grown for coffee-substitute production (Van Cutsem et al. 1993; Kiers et al. 2000; Doré and Varoquaux 2006). Ancient Witloof cultivars were open-pollinated populations forced in fields, whereas modern cultivars are generally F1 hybrids forced under a hydroponic system. These modern cultivars cannot be morphologically differentiated from each other, but only for their agronomical/production characteristics (Doré and Varoquaux 2006). Some crosses between Radicchio and Witloof gave rise to the red chicory forcing type (Van Kruistum 1997).

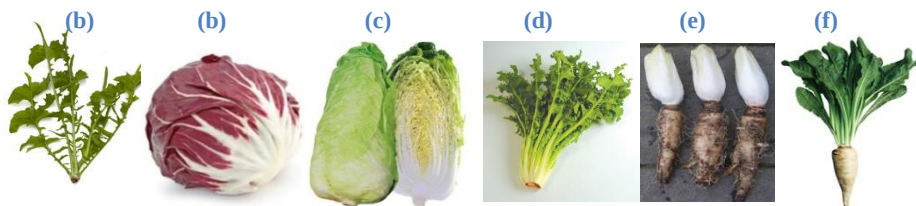


Figure 33: Chicories from different cultivar groups (a) wild individual (b) leaf chicory, Radicchio type (c) leaf chicory, Sugarloaf type, (d) leaf chicory, Catalogne type, (e) witloof, (f) industrial chicory.

Among wild forms of *C. intybus*, Kiaer et al. (2007) detected a degree of genetic differentiation from northern to southern Europe and also differentiation of two localized subgroups in Italy and Denmark.

Previously published genetic investigations of the *C. intybus* gene pool suffer from one or several of the following limitations. First, they sampled few individuals per cultivar (Kiers et al. 2000). Kiers (2000) provides the first genetic investigation of the overall cultivated chicory gene pool, which is an excellent starting point for anyone interested in the diversity of *C. intybus* and *Cichorium* as a whole. Second, these investigations analyzed few cultivars per cultivar group (Kiers et al. 2000, Kiaer et al. 2007). Third, some cultivar groups were not sampled (Baes and Van Cutsem 1993; Van Stallen et al. 2001). Finally, some studies were concerned with one cultivar group only (Bellamy et al. 1996; Koch and Jung 1997; De Proft et al. 2003; Van Cutsem et al. 2003).

The overall objective of our research is to investigate the genetic diversity and structure of the original germplasm used to breed the cultivars of industrial

chicory currently grown for inulin production. However, before investigating industrial chicory germplasm, it is first necessary to clarify the relationships of industrial chicory among *C. intybus* germplasm. As a consequence, the aim of the present study was to undertake a genetic investigation of cultivated *Cichorium* germplasm to overcome the above-mentioned limitations of previous studies. Therefore, the first step consisted of analyzing a panel of individuals representing most of the diversity present among cultivated *Cichorium* species: *C. endivia* individuals, wild *C. intybus* individuals, and individuals from several cultivars of all cultivar groups described in the Introduction. The ‘Magdeburg’ cultivar and some cultivars of the Catalogne subgroup were included in the study. To our knowledge, ‘Magdeburg’ and the Catalogne subgroup have not been included previously in a genetic investigation. ‘Magdeburg’ was included in the present study to evaluate the relationship between the Witloof group and its possible progenitor. Cultivars from the Catalogne subgroup were included to explore diversity among the widest range possible of leaf chicory cultivars. The second step of this study focused on a panel consisting of 1297 individuals from the 15 modern industrial chicory cultivars grown in the 1980s, as material for breeding of the current Belgian industrial chicory cultivar.

5.4 Materials and methods

5.4.1 Plant material

A collection of 19 *C. endivia* and 1413 *C. intybus* accessions was investigated. The *C. intybus* collection comprised 27 wild individuals from Europe, Iran, Pakistan, Russia, Turkey and the Syrian Arab Republic, 10 Sugarloaf, 12 Radicchio, 8 Catalogne, 42 Witloof and 1314 root chicory individuals, i.e. 17 individuals of ‘Magdeburg’ and 1297 individuals of 15 modern industrial chicory cultivars used to breed the current cultivars of industrial chicory (Table 3). ‘Magdeburg’ and the modern industrial cultivars were distinguished because the ‘Magdeburg’ cultivar arose much earlier (before 1850) than the modern cultivars (released after 1950) and also because the aim of the present study was to investigate genetic diversity among modern industrial cultivars. The first step was dedicated to the study of the genetic diversity among *C. endivia* and *C. intybus* based on a first panel that comprised three individuals of each of the 15 modern industrial cultivars and all accessions of all taxa, i.e. 190 individuals. The second step was dedicated to analysis of a second panel consisting of 1297 individuals of the 15 modern industrial cultivars.

Seeds of the modern industrial cultivars have been maintained at 4°C since 1990 by Chicoline®. All remaining accessions were obtained from the genebanks of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK; Gatersleben, Germany) and North Central Regional Plant Introduction Station

Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the
gene pool of industrial chicory

(NCRPIS; Ames, IA, USA), and certified seeds available from commercial suppliers (Aveve, Caillard, Cesens, Clause, Horticola, Semailles, Sementi, Sluis Garden, and Vilmorin). All individuals were grown in a greenhouse at Warcoing, Belgium between 2008 and 2013. Fresh leaf tissues were harvested at the two-leaf stage for DNA extraction.

Species	Cultivar group	Cultivar / Accession number	Ni	Breeding institution / seed distributor	Country	Year of release
<i>C. endivia</i>		CICH 1	2	IPK (**)	Greece	
<i>C. endivia</i>		CICH 14	2	IPK (**)	Greece	
<i>C. endivia</i>		CICH 15	2	IPK (**)	Greece	
<i>C. endivia</i>		CICH 50	3	IPK (**)		
<i>C. endivia</i>		CICH 86	3	IPK (**)		
<i>C. endivia</i>		CICH 87	1	IPK (**)		
<i>C. endivia</i>		Tosca	2	Vilmorin (*)	France	
<i>C. endivia</i>	Frisée	'Chicorée frisée Wallone' 'Despa'	1	AVEVE (**)	Belgium	
<i>C. endivia</i>	Scarole	'Chicorée scarole numéro 5'	3	AVEVE (**)	Belgium	
<i>C. intybus</i>	leaf (Catalogne)	'Catalogna gigante di chioggia'	3	Horticola	Italy	
<i>C. intybus</i>	leaf (Catalogne)	'Catalogna puntarelle a foglie strette'	3	Horticola	Italy	
<i>C. intybus</i>	leaf (Catalogne)	'Selvatica da campo'	1	Horticola	Italy	

Species	Cultivar group	Cultivar / Accession number	Ni	Breeding institution / seed distributor	Country	Year of release
<i>C. intybus</i>	leaf (Catalogne)	'Spadona'	1	Horticola	Italy	
<i>C. intybus</i>	leaf (Radicchio)	'Rossa di Verona'	3	Horticola	Italy	
<i>C. intybus</i>	leaf (Radicchio)	'Variegata di chioggia'	3	Sementi Cesena	Italy	
<i>C. intybus</i>	leaf (Radicchio)	'Carmen'	3	Clause (*)	France	
<i>C. intybus</i>	leaf (Radicchio)	'Rossa di Treviso'	3	Horticola	Italy	
<i>C. intybus</i>	leaf (Sugarloaf)	'Grumolo biondo'	1	Horticola	Italy	
<i>C. intybus</i>	leaf (Sugarloaf)	'Grumolo verde scuro'	3	Sementi Cesena	Italy	
<i>C. intybus</i>	leaf (Sugarloaf)	'Zuccherina di Trieste'	2	Horticola	Italy	
<i>C. intybus</i>	leaf (Sugarloaf)	'Bianca di Milano'	3	Horticola	Italy	
<i>C. intybus</i>	leaf (Sugarloaf)	'Pan di zucchero'	1	Horticola	Italy	
<i>C. intybus</i>	root	'Magdeburg' (CICH 792)	17	IPK (**)	Germany	
<i>C. intybus</i>	root	'Pévèle'	81	S.A. Florimond Desprez (*)	France	1984

Species	Cultivar group	Cultivar / Accession number	Ni	Breeding institution / seed distributor	Country	Year of release
<i>C. intybus</i>	root	'Fruitosa'	78	S.A. Florimond Desprez (*)	France	1984
<i>C. intybus</i>	root	'Cassel'	77	S.A. Florimond Desprez (*)	France	1986
<i>C. intybus</i>	root	'Orchies'	84	S.A. Florimond Desprez (*)	France	1987
<i>C. intybus</i>	root	'Tilda'	75	Bataille Alphonse (*)	France	1981
<i>C. intybus</i>	root	'Fredonia'	71	Sarea GmbH (*)	Austria	
<i>C. intybus</i>	root	'Polanowicka'	123	Centrum Obrotu Rolnego "Agro-Kal" sp. z.o.o. (*)	Poland	
<i>C. intybus</i>	root	'Kujawska'	103	Gospodarstwo Rolno-Nasienne KOW-MAR Maria J. Kowalska (*)	Poland	
<i>C. intybus</i>	root	'Dageraad'	116	Merelbeke (*)	Belgium	1981
<i>C. intybus</i>	root	'Halle'	120	Merelbeke (*)	Belgium	
<i>C. intybus</i>	root	'Sabau 3'	75	Ets Sabau (*)	France	

Species	Cultivar group	Cultivar / Accession number	Ni	Breeding institution / seed distributor	Country	Year of release
<i>C. intybus</i>	root	'Bergues'	72	S.A. Florimond Desprez (*)	France	1988
<i>C. intybus</i>	root	'Candi'	75	Selbat (*)	France	
<i>C. intybus</i>		'Novipa'	73	Merelbeke (*)	Belgium	1956
<i>C. intybus</i>	root	'Hicor'	74	Bataille Alphonse (*)	France	
<i>C. intybus</i>	Wild	PI 2006	1	NCRPIS (**)	Germany	
<i>C. intybus</i>	Wild	PI 652007	1	NCRPIS (**)	Poland	
<i>C. intybus</i>	Wild	PI 652008	1	NCRPIS (**)	Poland	
<i>C. intybus</i>	Wild	PI 652009	1	NCRPIS (**)	Poland	
<i>C. intybus</i>	Wild	PI 652019	1	NCRPIS (**)	Switzerland	
<i>C. intybus</i>	Wild	PI 652020	1	NCRPIS (**)	Hungary	
<i>C. intybus</i>	Wild	PI 652021	1	NCRPIS (**)	Hungary	
<i>C. intybus</i>	Wild	PI 652022	1	NCRPIS (**)	Hungary	

Species	Cultivar group	Cultivar / Accession number	Ni	Breeding institution / seed distributor	Country	Year of release
<i>C. intybus</i>	Wild	PI 652024	1	NCRPIS (**)	Germany	
<i>C. intybus</i>	Wild	PI 652025	1	NCRPIS (**)	Poland	
<i>C. intybus</i>	Wild	PI 652026	1	NCRPIS (**)	Iran	
<i>C. intybus</i>	Wild	PI 652027	1	NCRPIS (**)	Russia	
<i>C. intybus</i>	Wild	PI 652028	1	NCRPIS (**)	Russia	
<i>C. intybus</i>	Wild	Ames 24285	1	NCRPIS (**)	Pakistan	
<i>C. intybus</i>	Wild	PI 652030	1	NCRPIS (**)	Yugoslavia	
<i>C. intybus</i>	Wild	PI 652034	1	NCRPIS (**)	Poland	
<i>C. intybus</i>	Wild	Ames 26033	1	NCRPIS (**)	Portugal	
<i>C. intybus</i>	Wild	CICH100	3	IPK (**)	Iraq	
<i>C. intybus</i>	Wild	CICH192	3	IPK (**)	Syrian Arab Republic	
<i>C. intybus</i>	Wild	PI 269461	1	NCRPIS (**)	Pakistan	

Species	Cultivar group	Cultivar / Accession number	Ni	Breeding institution / seed distributor	Country	Year of release
<i>C. intybus</i>	Wild	PI 279705	1	NCRPIS (**)	Turkey	
<i>C. intybus</i>	Wild	PI 393818	1	NCRPIS (**)	Holland	
<i>C. intybus</i>	Wild	PI 531292	1	NCRPIS (**)	Hungary	
<i>C. intybus</i>	Witloof	'Mechelen'	10	Semailles (**)	Belgium	
<i>C. intybus</i>	Witloof	'Hollandse'	10	Semailles (**)	Belgium	
<i>C. intybus</i>	Witloof	'Mechelse'	10	Sluis Garden (**)	Holland	
<i>C. intybus</i>	Witloof	'Turbo' (F ₁ hybrid)	3	Caillard (**)	France	
<i>C. intybus</i>	Witloof	'Crenoline' (F ₁ hybrid)	3	Vilmorin (*)	France	
<i>C. intybus</i>	Witloof	'Zoom' (F ₁ hybrid)	3	Vilmorin (**)	France	
<i>C. intybus</i>	Witloof	'Flash' (F ₁ hybrid)	3	Sluis Garden (**)	Holland	

Table 3: Accessions included in the study (first and second panel); for each cultivar or accession the following information is provided: the cultivar group (the subgroup is specified in parentheses); cultivar name or accession identifier; number of individuals per accession or per variety (*Ni*); breeding institution (*) or seed distributor (**); country of origin or country in which the breeding institution or seed distributor is located; year of release of the variety.

5.4.2 DNA extraction and SSR analyses

Genomic DNA was extracted in accordance with the protocol described by Muys et al. (2013). All individuals were genotyped with 15 SSR markers. The SSRs were identified by selective hybridization (Hamilton et al. 1999) and the genotyping of all individuals was performed by Cosucra Belgium (proprietary sequences). The PCR amplifications were performed in accordance with the protocol described by Muys et al. (2013). Primer pairs are available upon request for academic use only. The linkage groups in which 11 of the 15 SSRs are located were identified by Muys et al. (2013) (Table 2).

The informativeness of the SSR markers was evaluated by the number of alleles, the polymorphism index content (PIC, defined as $PIC = 1 - \sum p_i^2$, where p_i is the frequency of the i th allele of the locus), and the repetition range. These variables and the proportion of missing data were computed with R 3.0.0 software (R Development Core Team 2013).

5.4.3 Population analyses

The structure among and within the different cultivars was evaluated via cluster and multivariate (principal component analysis; PCA) analyses, the assignment method, and analysis of molecular variance (AMOVA).

Cluster analysis was performed using the unweighted pair group method with

arithmetic mean (UPGMA) method, at the individual level with the Dps distance ($Dps = \exp[1 - \text{proportion of shared alleles}]$) (Bowcock et al. 1994), and at the cultivar level with F_{st} (Nei 1973) and D_{st} (Nei's standard distance; Nei 1972). At the cultivar level, a bootstrap analysis with 1000 replicates was performed to evaluate the statistical support for each node. The cluster analysis and PCA were performed with R 2.14.0 software (R Development Core Team 2013), using functions in the packages adegenet (Jombart 2008), ape (Paradis et al. 2004), and ade4 (Chessel et al. 2004).

The assignment method was applied with the software Structure (Prichard et al. 2000) using the admixture model. Different numbers of putative populations (K) were tested, ranging from 2 to 12. Ten independent runs were carried out per K and 10^6 Markov chain Monte Carlo iterations were used after a burn-in of 10^5 steps. The best K was evaluated in accordance with the Evanno method (Evanno et al. 2005). The simulations with the highest $L(K)$ are presented in the figures. AMOVA analyses and F_{st} values were computed with Arlequin version 3.11 software (Excoffier and Lischer 2010). The genetic diversity of the different cultivars was evaluated as the number of alleles, the number of private alleles (alleles present into one group only), and Nei's diversity index (I_{Nei}), computed as $I_{Nei} = 1 - (\sum \sum p_{ij}^2)/n$, where p_{ij} is the frequency of the i th allele of the j th locus (with a total number of n loci).

5.5 Results

The results are divided into three sections. The first section describes the characteristics of the SSR markers for all available individuals (first and second panel). The second section describes the results concerning the first step of the study, i.e. the population structure within the two cultivated *Cichorium* species (first panel). The third section describes the results concerning the second step, i.e. the population structure of the modern industrial cultivars (second panel), and includes all available individuals of the modern industrial cultivars.

5.5.1 SSR polymorphism (first and second panel)

The PIC values and allelic richness of the SSR markers were high and variable, ranging from 0.64 to 0.92 and 8 to 55 alleles per locus, respectively (Table 4). The F₁ Witloof cultivars ‘Créoline’, ‘Turbo’, and ‘Zoom’ showed variation between individuals at one, two and five loci, respectively. ‘Flash’ did not show any polymorphism.

Loci	Na	PIC	Repetition range	Linkage group
GA007	8	0.67	211–230	unknown
GA036	29	0.87	163–214	3
GA118	47	0.82	103–218	3
GA187	26	0.81	124–186	8
GA014	20	0.72	113–152	6
GA019	46	0.87	298–362	2
GA075	40	0.87	188–266	5
GA304	30	0.64	201–252	unknown
GA355	50	0.91	166–266	1
GA386	20	0.79	121–176	?
GA495	55	0.92	177–264	6
GA519	27	0.79	140–193	7
GA571	22	0.74	194–235	7
GA357	48	0.79	161–256	2
GA395	26	0.82	193–247	unknown
Mean	32.93	0.8		
Total	494			

Table 4: Number of alleles (Na), polymorphism index content (PIC), repetition range and linkage group

Overall, there was 0.9% missing data. Nevertheless, the missing data were unequally distributed. Few amplification products were observed at the locus GA187 in the Radicchio subgroup (87.5% missing data) and at five loci (GA007, GA036, GA187, GA075, and GA357; 79–100% missing data) for the *C. endivia* individuals and seven wild *C. intybus* individuals from Iraq (three individuals),

Pakistan (one individual) and the Syrian Arab Republic (three individuals). Two SSR analyses with DNA from two independent extractions were performed on those individuals, which confirmed the unsuccessful amplification at these five loci. This finding was a strong indication of genetic differentiation between, on the one hand, *C. endivia* and the seven wild *C. intybus* and, on the other hand, the cultivated *C. intybus*. The missing data of these five loci in the *C. endivia* individuals and in the seven previously mentioned wild *C. intybus* individuals were analyzed both as missing data and null allele. The results presented the section “Genetic diversity among cultivated *Cichorium*” were obtained with missing amplification coded as null allele.

The number of alleles, private alleles, and I_{Nei} were extremely high in the population of wild *C. intybus* individuals ($I_{Nei} = 0.95$), high in the modern industrial chicory and leaf chicory group ($I_{Nei} = 0.90\text{--}0.91$), intermediate in *C. endivia* ($I_{Nei} = 0.81$), and extremely low in the Witloof group ($I_{Nei} = 0.22$) (Table 5). The number of alleles, private alleles, and I_{Nei} were high in the industrial chicory group. However the number of modern industrial individuals was 45 to 75 times higher than those of the other groups, which is likely to have affected the allelic diversity results.

Group	Individual number	Allele number	Private allele number	I_{Nei}
<i>C. endivia</i>	19	61	11	0.81
<i>C. intybus</i>	182	300	249	0.91
Modern industrial cultivars (first panel)	66	93	14	0.95
Modern industrial cultivars (second panel)	1297	155	26	0.9
‘Magdeburg’	17	78	4	0.65
Witloof	42	43	3	0.22
Wild	27	220	78	0.95
Leaf cultivars	30	138	32	0.91
Radicchio	12	75	9	0.76
Sugarloaf	10	71	8	0.73
Catalogne	8	56	9	0.8

Table 5: Individual number, allele number, private allele number, and Nei’s genetic index (I_{Nei}) of the different groups of *C. intybus* and *C. endivia* accessions.

The allelic patterns among the modern industrial cultivars (second panel) showed notable variability (Table 6). The total number of alleles and the number of private alleles per cultivar ranged from 92 to 173 and from 0 to 7, respectively. The Nei’s genetic index value was moderately high, ranging from 0.57 to 0.77.

Cultivar	Individual number	Allele number	Private allele number	I_{Nei}
‘Pévèle’	81	143	4	0.72
‘Fruitosa’	78	149	1	0.77
‘Cassel’	77	152	5	0.73
‘Orchies’	84	152	4	0.77
‘Tilda’	75	133	2	0.67
‘Fredonia’	71	120	2	0.70
‘Polanowicka’	123	173	7	0.71
‘Kujawska’	103	146	7	0.76
‘Dageraad’	116	103	3	0.60
‘Halle’	120	92	4	0.57
‘Sabau 3’	75	136	3	0.72
‘Bergues’	72	136	1	0.74
‘Candi’	75	139	1	0.70
‘Novipa’	73	102	0	0.62
‘Hicor’	74	115	1	0.64

Table 6: Individual number, allele number, private allele number, and Nei’s genetic index (I_{Nei}) of 15 cultivars of modern industrial chicory (second panel).

5.5.2 Genetic diversity among cultivated Cichorium (first panel)

Population assignment was tested for K ranging from 2 to 13. The Evanno method showed that the most relevant K was 4. Nevertheless, clustering performed with K up to 8 produced coherent results because the putative

populations identified corresponded to the different cultivar groups, and each individual was mostly assigned to one putative population only (Figure 34). For $K = 4$, one putative population included the Witloof cultivars, one comprised the root cultivars, one contained most of the wild *C. intybus* individuals and the leaf cultivar individuals and the last one comprised the accessions for which no amplification products were obtained in five loci (see materials and methods), *i.e.* *C. endivia* and seven wild *C. intybus* individuals from Iraq, Pakistan and the Syrian Arab Republic. For $K = 8$, each putative populations corresponded to the same genetic groups identified in the cluster analysis: (i) Witloof, (ii) ‘Magdeburg’, (iii) modern industrial cultivars, (iv) Sugarloaf, (v) Radicchio, (vi) Catalogne, (vii) most of the wild *C. intybus* individuals, and (viii) the *C. endivia* and the seven wild *C. intybus* individuals mentioned above.

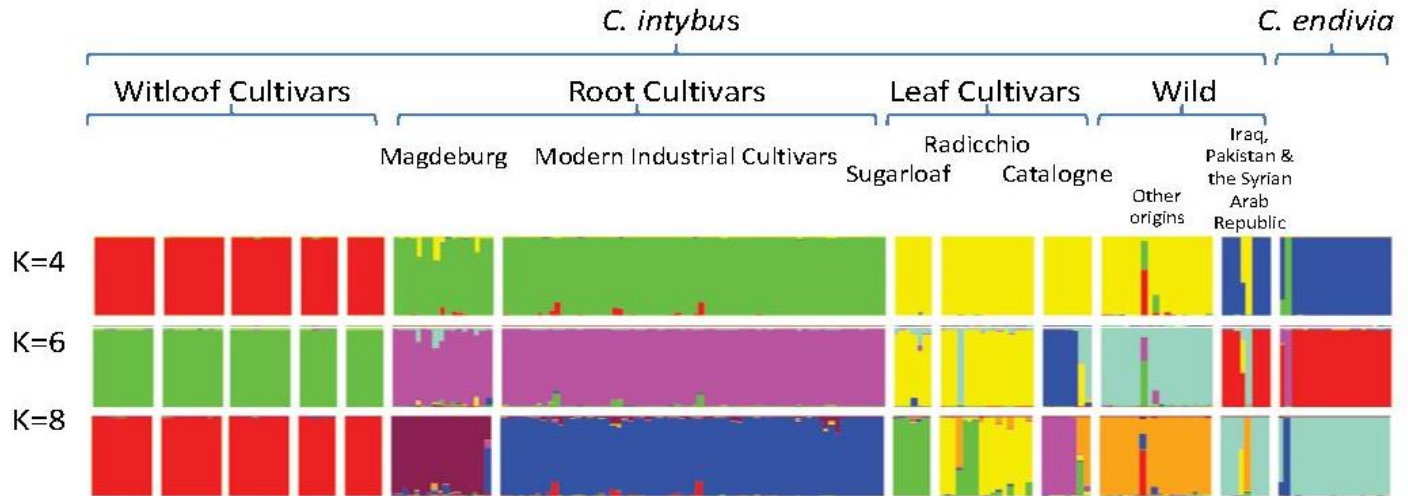


Figure 34: Genetic structure results for $K = 4, 6$, and 8 for *C. endivia* and the *C. intybus* cultivar groups (first panel).

The UPGMA dendrogram based on the Dps distance for the 15 loci in individuals of the two cultivated species divided the panel in two clusters (Figure 35). The first one included most of the *C. intybus* and the second included all *C. endivia* individuals and seven wild *C. intybus* individuals from Iraq, Pakistan and the Syrian Arab Republic. The latter seven individuals were those for which amplification of five loci was unsuccessful. Most of the other wild individuals were grouped in two distinct clusters within the *C. intybus* cluster. The six remaining wild individuals were placed randomly within the *C. intybus* cluster. Each subcluster of the *C. intybus* cluster corresponded to a specific cultivar group: one to the root cultivars, one to the Witloof group, one to the leaf cultivars, and two distinct clusters including most of the wild individuals, as mentioned above. The root cultivar and Witloof clusters were linked, although with a high dissimilarity. The root cultivar cluster was clearly divided into two distinct clusters: one cluster contained all ‘Magdeburg’ individuals and the other cluster included all but one individual of the modern industrial chicory cultivars used for selection of the current industrial cultivars. Within the Witloof cluster, each cultivar population formed a distinct subcluster, except for three individuals of ‘Mechelen’ that clustered within the ‘Hollandse’ cluster (data not shown). With regard to the F1 cultivars, three distinct clusters were formed: one corresponded to ‘Crénoline’, one to ‘Zoom’, and one combined the ‘Flash’ and ‘Turbo’ individuals. The leaf cultivars and one wild individuals cluster were linked in the same cluster. The leaf cultivar cluster was divided into three distinct subclusters. One cluster corresponded to the Catalogne subgroup, one to the Radicchio subgroup, and one to the Sugarloaf subgroup. The Radicchio and Sugarloaf subgroups were linked. Two-leaf cultivar individuals and several small clusters of wild individuals were linked at other locations in the dendrogram. The

Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the
gene pool of industrial chicory

PCA results showed a very similar population structure to that of the cluster analysis (Supplementary material 1).

Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the gene pool of industrial chicory

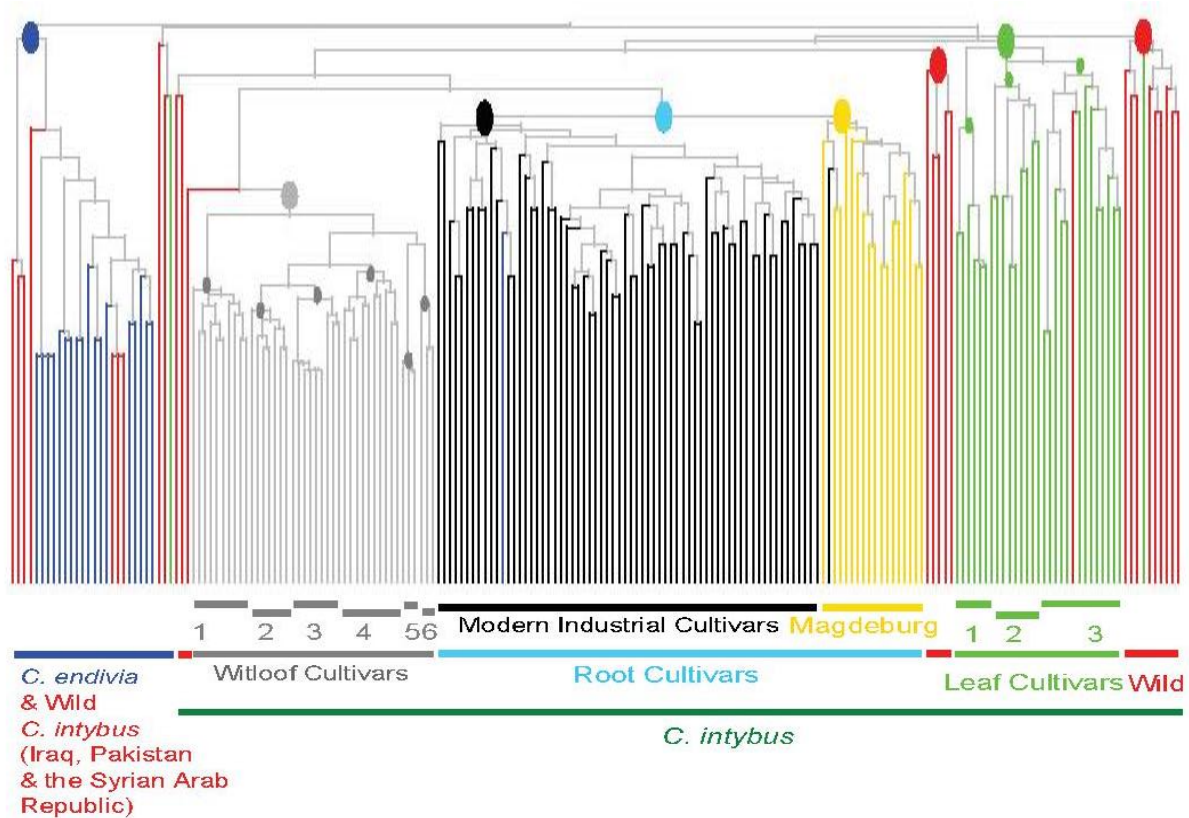


Figure 35: UPGMA dendrogram based on Dps distances among *C. endivia* and the *C. intybus* cultivar groups and subgroups (first panel). The branch color corresponds to the cultivar group. The dots correspond to the last junction including all individuals from one cluster. For the leaf cultivars, groups 1, 2, and 3 correspond to the Catalogne, Sugarloaf, and Radicchio subgroups, respectively. For the Witloof cultivars, the groups 1, 2, 3, 4, 5, and 6 correspond to the cultivars ‘Mechelse’, ‘Mechelen’, ‘Flash’ and ‘Turbo’, ‘Hollandse’, ‘Créneline’ and ‘Zoom’, respectively. The AMOVA analyses showed that segregation in two species (*C. intybus* vs *C. endivia*) explained a small fraction of the genetic variation of our panel (22.2%), whereas the variation within *C. intybus* (root cultivars vs Witloof vs leaf cultivars vs wild *C. intybus* vs seven wild *C. intybus* individuals from Iraq, Pakistan and the Syrian Arab Republic) represented 19.64% of the genetic variation (Table 7). Nevertheless, the *C. endivia* and *C. intybus* sample sizes were highly unbalanced, which could have affected the results. The AMOVA applied to the different cultivar groups of cultivated *C. intybus* revealed only that the cluster of root cultivars vs Witloof vs leaf cultivars explained 20.39% of the genetic variation. The separation of the leaf cultivars into the three cultivar subgroups explained 12.74% of the genetic variation within the leaf cultivar group. The F_{st} between *C. endivia* and the different *C. intybus* groups or subgroups ranged from 0.27 to 0.47, which is a bit higher than between *C. endivia* and all *C. intybus* cultivars together (Table 8).

Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the gene pool of industrial chicory

	<i>C. endivia</i> vs <i>C. intybus</i>	root cultivars vs Witloof vs leaf cultivars vs wild <i>C. intybus</i> vs seven wild <i>C.</i> <i>intybus</i> individuals from Iraq, Pakistan and the Syrian Arab Republic	Root cultivars vs Witloof vs leaf cultivar	Radicchio vs Sugarloaf vs Catalogne
Among groups	22.23	19.64	20.39	12.74
Within groups	77.7	80.36	90.32	87.26

Table 7: AMOVA results for the first panel .The column headings describe the groups used in the AMOVA.

F_{st}	C. <i>endivia</i> a	Modern industrial cultivars	'Magdeburg'	Witloof	Wild	Wild western Asia	Radicchio	Sugarloaf
<i>C. intybus</i>	0.22							
Modern	0.305	0						
industrial								
cultivars								
'Magdeburg'	0.308	0.081	0					
Witloof	0.47	0.276	0.28	0				
Wild*	0.284	0.10	0.085	0.271	0			
Wild	0.191	0.352	0.358	0.54	0.28	0		
western Asia								
Leaf cultivar	0.273	0.108	0.120	0.326	0.054	0.311		
Radicchio	0.33	0.12	0.14	0.37	0.07	0.38	0	
Sugarloaf	0.33	0.17	0.18	0.40	0.09	0.39	0.11	0
Catalogne	0.33	0.14	0.15	0.39	0.08	0.38	0.11	0.12

Table 8: Fixation index (F_{st}) values between the groups of *C. intybus* accessions and *C. endivia* (first panel). Wild*: the wild *C. intybus* except from those coming from western Asia (Pakistan, Syria and Iraq) Wild*: wild *C. intybus* accessions except for those originating from western Asia (Pakistan, Syrian Arab Republic and Iraq).

5.5.3 Genetic diversity among modern industrial cultivars (second panel)

The assignment method was tested for K ranging from 2 to 13. The Evanno method showed that the most relevant K was 3, but higher levels of clustering were coherent with the cultivars (Figure 36). For $K = 3$, most of the individuals were clearly assigned to one of the putative populations so that three groups were distinguishable: one group included all of the French cultivars ('Pévèle', 'Orchies', 'Fruitosa', 'Cassel', 'Tilda', 'Sabau 3', 'Bergues', 'Candi' and 'Hicor'), a second group included the Belgian cultivars 'Dageraad' and 'Halle', and a third group combined the cultivars 'Fredonia', 'Polanowicka', 'Kujawska' and 'Novipa'. At this population structure level, most individuals of 'Fredonia' had a high proportion of their genome assigned to the two other putative populations (60% of the individuals had more than 20% of their genome assigned to the two other populations). At $K = 7$, each cultivar of the third group ('Fredonia', 'Polanowicka', 'Kujawska' and 'Novipa') was mainly assigned to one putative population and the first group (French cultivars) was divided into two subgroups: (a) 'Pévèle', 'Orchies', 'Fruitosa', and 'Cassel', and (b) 'Tilda', 'Sabau 3', 'Bergues', 'Candi' and 'Hicor'. The individuals of subgroup (b) were partially assigned to other populations. At $K = 11$, the cultivars of the first group (French cultivars) were distributed among five putative populations as follows: the cultivars (i) 'Pévèle' and (ii) 'Cassel' were mainly assigned to one population each, whereas the cultivar pairings (iii) 'Tilda' and 'Hicor' and (iv) 'Orchie' and

Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the
gene pool of industrial chicory

‘Fruitosa’ were grouped into two different populations, and (v) the cultivars
‘Sabau 3’, ‘Bergues’ and ‘Candi’ were mainly assigned to one population.

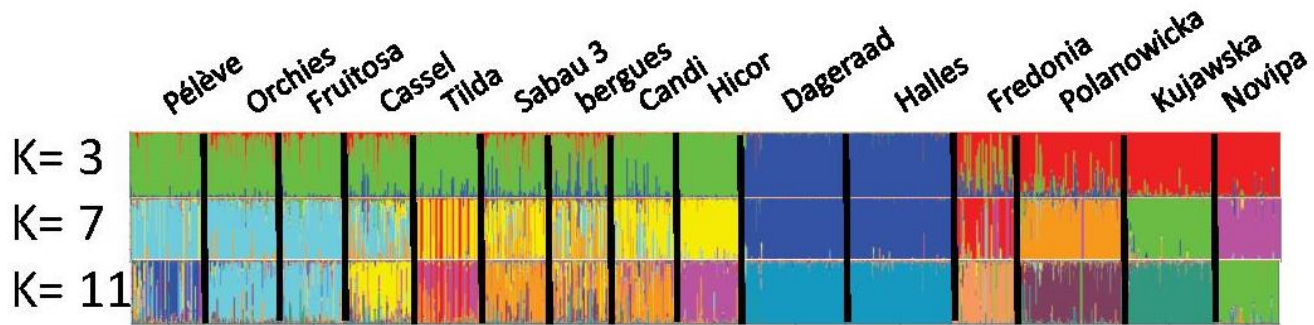


Figure 36: Genetic structure results for $K = 3, 7$, and 11 for the modern industrial chicory cultivars (second panel).

The UPGMA dendrogram based on the Dps distance separated the cultivars into different clusters. Two clusters contained most individuals from all the French cultivars: ‘Pévèle’, ‘Orchies’, ‘Fruitosa’, ‘Cassel’, ‘Tilda’, ‘Sabau 3’, ‘Bergues’, ‘Candi’ and ‘Hicor’ (Figure 37). Individuals of the cultivars ‘Tilda’, ‘Sabau 3’, ‘Bergues’ and ‘Candi’ were placed in the two clusters, whereas individuals of all other cultivars were only placed in one cluster. No internal structure was evident within the two French cultivar clusters. The two Belgian cultivars ‘Dageraad’ and ‘Halle’ were combined in one cluster. The individuals of ‘Polanowicka’ (Poland), ‘Kujawska’ (Poland) and ‘Novipa’ (Belgium) were present in three different clusters. The individuals of ‘Fredonia’ (Austria) were placed in different but closely linked small clusters. The cluster including the ‘Novipa’ individuals was well differentiated from the other clusters. Nevertheless, no clear structure among the other clusters described above was discernible within this dendrogram. Two dendrograms based on the allelic frequencies of the cultivars were generated (data not shown). In both cases, three clusters were retrieved. One cluster contained the French cultivars, and a second cluster contained the cultivars ‘Novipa’, ‘Polanowicka’, ‘Kujawska’ and ‘Fredonia’. The clustering of the cultivars within these two clusters changed with the distance measure used and received low bootstrap support. The third cluster contained the cultivars ‘Dageraad’ and ‘Halle’ and was supported by high bootstrap values. The PCA performed on the same data produced very similar results (Supplementary material 2).

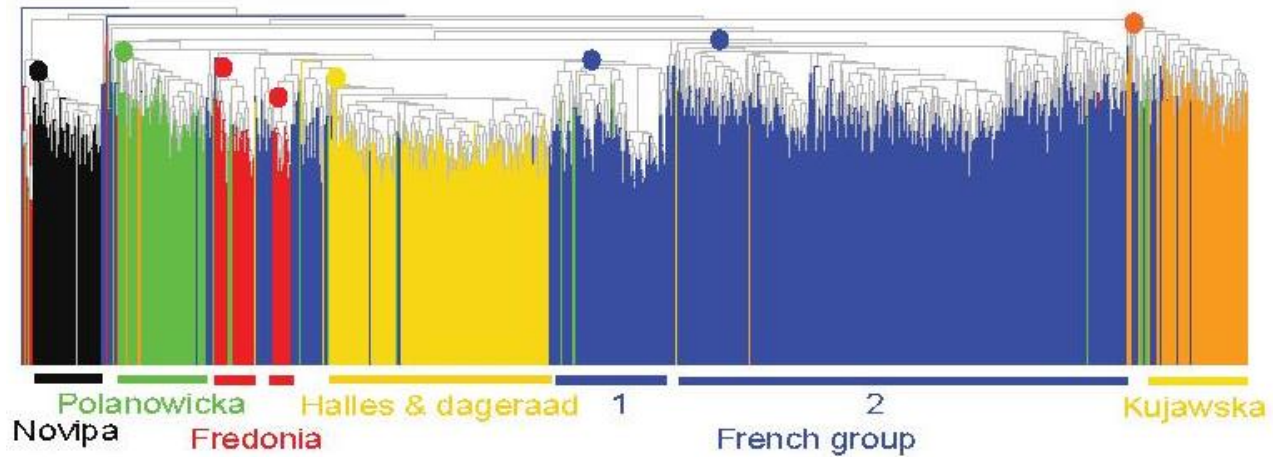


Figure 37: UPGMA dendrogram based on Dps distances among modern industrial chicory cultivars (second panel). The branch color corresponds to the cultivar group. The dots correspond to the last junction including all individuals from one cluster.

The AMOVA analyses showed that most of the variation (87–88%) occurred within the cultivars. The segregation of the cultivars obtained with the assignment method for $K = 3$ and 7 respectively explained 6.17% and 8.43% of the genetic variation (Table 9). Segregation into 11 groups did not notably increase the proportion of variance explained by the different groups (8.76%).

	$K = 3$	$K = 7$	$K = 11$
Among groups	6.17	8.43	8.76
Among cultivars within groups	6.92	3.75	2.81
Within cultivars	86.91	87.82	88.83

Table 9: AMOVA results for three levels of cultivar groupings obtained with the assignment method for $K = 3$, 7 and 11 for the modern root chicory cultivars (second panel).

5.6 Discussion

5.6.1 Species delimitation

The different analyses confirmed the genetic differentiation of *C. endivia* and *C. intybus* (Vermeulen et al. 1994; Kiers et al. 1999, 2000). Firstly, five of the 15 SSR loci were not amplified in *C. endivia*. Secondly, the assignment, clustering and multivariate analysis clearly separated *C. endivia* from *C. intybus*. The F_{st} between *C. endivia* and *C. intybus* ($F_{st} = 0.22$) was very close to the value between the cluster including *C. intybus/C. spinosum* and the cluster including *C. pumilum/C. endivia* ($F_{st} = 0.23$) estimated by Gemeinholzer and Bachmann (2005).

The unsuccessful amplification of five of the 15 SSR loci in *C. endivia* is a strong indicator of the differentiation between both species but also leads to some analytical problems. Indeed the absence of amplification can be coded as missing data or as null allele. Both methods involve some bias. When the absence of amplification is coded as missing data, the dataset do not integrate some important information concerning the differentiation between both species and, hence, underestimates the differentiation between them. When missing amplification are coded as null allele, it artificially increases the homogeneity within *C. endivia*, which may artificially increase the observed differentiation between species in the different analysis. However, we realized all population analyses with both methods and they led to the same conclusions. The

segregation between the two species was more important in the null allele method than in the missing data method and the population structure within *C. intybus* was very similar for both methods. We presented the result obtained by the null allele method because they were the clearest ones.

Our dataset did not identify any genetic structure within *C. endivia*, which is consistent with the results of Kiers et al. (2000)

5.6.2 Relationships among the three *C. intybus* cultivar groups

The present analyses confirmed the genetic differentiation and the constitution of genetic groups within the cultivated germplasm of *C. intybus* (Witloof, root cultivars, and leaf cultivars). Indeed, the majority of the individuals of each cultivar group clustered together both in the cluster and assignment analyses. The Witloof group was closer to the root cultivars than to the leaf cultivars and the wild individuals. The leaf cultivars and wild *C. intybus* individuals appeared to be closely related with each other.

The differentiation between cultivated forms compared to wild individual was clear in the different analyses. Moreover, PCA results showed a clear signature of the divergent selection of the leaf group compared to the root and Witloof groups: the leaf group was situated on one side of the wild individuals group whereas the root and Witloof groups were situated on the other side. The branching structure of UPGMA tree could not reflect this complex breeding

history and wild individuals clustered in different small cluster between the cluster including leaf group and the cluster including the root and Witloof groups.

The genetic diversity was extremely high among the wild individuals, high in the leaf cultivar group and in the root cultivar group, and extremely low in the Witloof group (I_{Nei} of 0.95, 0.91, 0.90, and 0.22, respectively). This finding could indicate that, in contrast to the Witloof group, the leaf and root cultivars have not undergone intensive selection or a marked bottleneck during selection. Baes et al. (1995) observed that genetic diversity was higher in the leaf cultivar group than in the root cultivar group. However, these authors based their conclusions on investigation of three loci only and thus are likely to be less accurate than ours, which are based on 15 loci.

5.6.3 ‘Magdeburg’, Witloof and modern industrial cultivars

‘Magdeburg’ is an ancient root cultivar mainly used as a coffee substitute (Doré and Varoquaux 2006). It is known to have been the source of the Witloof selection around 1850, which is the reason that this cultivar was included in the present study. All analyses revealed that the root cultivars and Witloof are genetically well-differentiated groups, but they formed a homogeneous group compared with the leaf cultivars and the wild *C. intybus* individuals. ‘Magdeburg’ and the modern industrial cultivars are closely related but genetically distinct groups. Considering that ‘Magdeburg’ was the original

breeding material for the Witloof group, a closer genetic relationship between ‘Magdeburg’ and Witloof cultivars might have been expected.

The close relationship between ‘Magdeburg’ and the modern industrial cultivars reflects the importance of ‘Magdeburg’ and other similar ancient coffee-substitute cultivars in the breeding history of the modern industrial cultivars. Compared with the Witloof group, the modern industrial cultivars contained a larger genetic base, had undergone less intensive selection, and no F_1 cultivars have been raised. Therefore, the modern industrial cultivars showed high genetic diversity and intermediate differentiation from ‘Magdeburg’, whereas the Witloof group showed low genetic diversity and strong differentiation from ‘Magdeburg’ ($F_{st} = 0.28$) and from all other *C. intybus* groups and subgroups (F_{st} ranging from 0.27 to 0.40).

Interestingly, although assignment, cluster and PCA analyses indicated that ‘Magdeburg’ was genetically closer to the modern industrial cultivars than to the Witloof group, the F_{st} values revealed that ‘Magdeburg’ was less clearly differentiated from wild *C. intybus* individuals than from the modern industrial cultivars ($F_{st \text{ wild/Magdeburg}} = 0.08$, $F_{st \text{ wild/industrial}} = 0.10$). This may reflect the shorter breeding history of ‘Magdeburg’ (raised before 1850) compared with the modern industrial cultivars analyzed here, which were raised more than 100 years later. The smaller I_{Nei} of ‘Magdeburg’ compared with that of the modern industrial cultivars may be explained by the lower number of individuals analyzed (17 vs 1168 individuals, respectively).

5.6.4 Leaf cultivars

The assignment, cluster and PCA analyses clearly divided the leaf cultivars into three subgroups: the Radicchio, the Sugarloaf and the Catalogne. This confirms the genetic differentiation between Radicchio and Sugarloaf described by Kiers et al. (2000). The present study also included representatives of the Catalogne subgroup, which consists of leaf chicory cultivars characterized by long thin leaves and the absence of a compact heart. To our knowledge, this subgroup has never been included in any genetic diversity study. The present study revealed that this subgroup, in addition to its phenotypic features, shows clear genetic differentiation from both the Radicchio and the Sugarloaf subgroups. Some breeders and seed producers designate the Catalogne cultivars as a wild or wild-like chicory type because this subgroup is phenotypically intermediate between wild chicory and the other leaf cultivars. Nevertheless, the Catalogne subgroup clustered with the other cultivated leaf cultivars and not with the wild *C. intybus* individuals, therefore designating them wild or wild-like chicory is inappropriate.

5.6.5 Wild individuals

For seven wild *C. intybus* individuals from Iraq, Pakistan, and the Syrian Arab Republic no amplification product was obtained for the same five loci as for *C. endivia*. This finding, along with the assignment, clustering, PCA and F_{st} analysis revealed that these seven *C. intybus* individuals were genetically distinct from the other wild *C. intybus* accessions (from Europe, Iran, Pakistan, Russia, and

Turkey) but genetically closer to *C. endivia* than to the other *C. intybus* accessions. Gemeinholzer and Bachmann (2005) also observed differentiation between wild *C. intybus* from Europe and some individuals from Central Asia (Armenia and Georgia) ($F_{st} = 0.24$). In their study, the wild accessions from Armenia and Georgia were clustered with *C. spinosum*, whereas the *C. endivia* accessions clustered with *C. pumilum*. Therefore, the analyses of Gemeinholzer and Bachmann (2005) clearly distinguished *C. endivia* from wild *C. intybus* from Armenia and Georgia. We cannot exclude the possibility that the seven individuals from Iraq, Pakistan, and the Syrian Arab Republic might have been misidentified because those individuals were collected within the distribution area of *C. calvum* and *C. pumilum*. The present study does not reject or confirm the close relationship of *C. endivia* with *C. calvum* and *C. pumilum*, but highlights the differentiation of *C. endivia* and some wild *C. intybus* accessions collected in western Asia (Iraq, the Syrian Arab Republic) or a neighboring region (Pakistan) from the other *C. intybus* provenances.

5.6.6 Modern industrial cultivars

In the second part of our study, we analyzed a collection of 15 cultivars of modern industrial chicory used to breed the current industrial chicory cultivars grown for inulin production. Thus, they are a snapshot of the original genetic diversity of the modern industrial cultivars. As observed within the first panel, this second panel was characterized by a high level of genetic diversity. ‘Magdeburg’ was not included in this part of the study because it represents a

distinct gene pool and was not directly included in the breeding of the modern industrial cultivars (at least in Belgium).

The assignment, clustering and PCA analyses showed that (i) the nine French cultivars formed one group, (ii) the Belgian cultivars ‘Dageraad’ and ‘Halle’ formed a uniform and genetically distinct group, (iii) ‘Novipa’ (Belgium), ‘Polanowicka’ (Poland), ‘Kujawska’ (Poland) and ‘Fredonia’ (Austria) were clearly differentiated from each other, (iv) ‘Fredonia’ showed some similarities with the French group, and (v) the relationships among the identified cultivars group were not clear. Most of the genetic diversity ($\pm 87\%$) was observed within the cultivars, as expected for open-pollinated population cultivars. Segregation into seven cultivar groups ($K = 7$) explained most of the genetic variability among the cultivars. These results are in agreement with those of Koch and Jung (1997) concerning the high genetic variability within the root chicory germplasm and the low differentiation between cultivars. However, the SSR markers were able to identify clearly groupings of some cultivars, whereas the RAPD and AFLP markers used by Koch and Jung did not. The high intra- and inter-cultivar genetic diversity is also associated with high phenotypic variability in root and leaf size and shape characteristics and in carbohydrate characteristics (Raulier unpublished results).

The cluster and PCA analyses were unable to resolve genetic structure within the French cultivars group, but the assignment method divided the French group into 2 or 5 subgroups (for $K = 7$ and 11, respectively). The segregation of the French cultivars into subgroups explained a minor proportion of the genetic variability within the French group (2% and 2.5% for two and five subgroups, respectively).

The main conclusion is that different French breeding institutions have used a common gene pool that has led to weakly genetically differentiated cultivars. Although these cultivars showed low genetic differentiation, some cultivars show significant differences in root and leaf size and shape characteristics and in carbohydrate characteristics (Raulier unpublished results).

The cultivars that were not bred by French institutions were more strongly genetically differentiated from each other and from the French group. These cultivars were bred by institutions located in Belgium ('Novipa', 'Dageraad' and 'Halle') or in eastern European countries ('Fredonia', 'Polanowicka' and 'Kujawska'). All of the Belgian cultivars sampled here were raised by Merelbeke. 'Novipa', released in 1956, is the oldest cultivar included in the present study and is characterized by a long taproot, typical of the coffee-substitute chicory cultivars. The cultivars 'Dageraad' and 'Halle' were released in the early 1980s. Both cultivars produce a taproot typical of industrial chicory. Those two cultivars appeared to belong to the gene pool. The cultivars raised in eastern Europe were clearly differentiated from each other.

5.7 Conclusion and perspectives

For the first section of the present study, the combination of co-dominant markers and a high number of individuals allowed improved discriminatory power over that of the study by Kiers et al. (2000) and confirmed the genetic differentiation of the different cultivar groups and subgroups. The inclusion of *C.*

endivia, wild *C. intybus* individuals (in particular from western Asia), and the Catalogne cultivars highlights the complex breeding history of these two species.

The modern industrial cultivars that were used for the recent breeding of cultivars for inulin production show high genetic and phenotypic variability. Therefore, this collection represents a useful gene pool for the breeding purpose, i.e., for the identification and characterization of genes of interest responsible for important traits (e.g., yield, inulin quality, and disease resistance) or for the introgression of these traits in the breeding material. A sample of this panel was used by Dauchot et al. (2014) to identify polymorphism among fructan exohydrolase genes associated with enhanced resistance to post-harvest inulin depolymerization. Given that their phenotype is not very different from that of the current cultivars, the introgression of interesting characters from the modern cultivar should be manageable, even for open-pollinated cultivars such as industrial chicory.

Acknowledgments The authors acknowledge the Walloon Region (DGARNE-Belgium) for supporting this research (grant D31-1221).

5.8 References

- Baes PG, Van Cutsem PJ (1993) Isozyme polymorphism in three gene pools of cultivated chicory (*Cichorium intybus* L.). *Euphytica* 71:143–150
- Bellamy A, Vedel F, Bannerot H (1996) Varietal identification in *Cichorium intybus* L. and determination of genetic purity of F₁ hybrid seed samples, based on RAPD markers. *Plant Breed* 115:128–132
- Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 68:455–457
- Chessel D, Dufour AB, Thioulouse J (2004) The ade4 package - I: One-table methods. *R News* 4:5–10
- Dauchot N, Raulier P, Maudoux O, Notté C, Bertin P, Draye X, Van Cutsem P (2014) Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest inulin depolymerization. *Theor Appl Genet* 127:125–135
- De Proft M, Van Stallen N, Veerle N (2003) Breeding and cultivar identification of *Cichorium intybus* L. var. *foliosum* Hegi. In: Van Hintum ThJL, Lebeda A, Pink D & Schut JW (eds) *Eucarpia Leafy Vegetables*. Centre for Genetic Resources, Wageningen, pp 83–90
- Doré C, Varoquaux F (2006) Histoire et amélioration de cinquante plantes cultivées. Institut national de la recherche agronomique, Paris
- Evanno G, Regnaut S, Goudet JG (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14:2611–2620

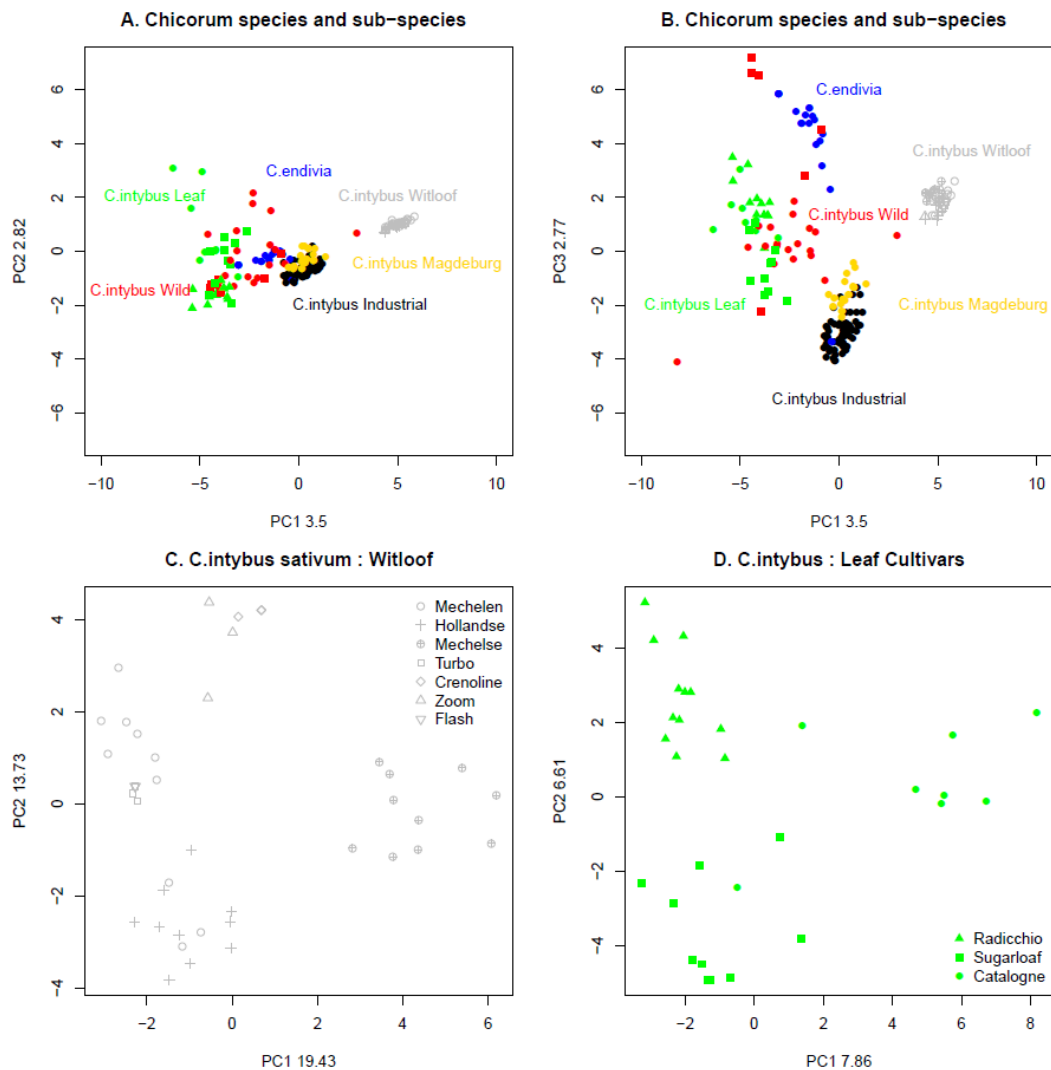
- Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10:564–567
- Gemeinholzer B, Bachmann K (2005) Examining morphological and molecular diagnostic character states of *Cichorium intybus* L. (Asteraceae) and *C. spinosum* L. *Plant Syst Evol* 253:105–123
- Hamilton MB, Pincus EL, Di Fiore A, Fleischer RC (1999) Universal linker and ligation procedures for construction of genomic DNA libraries enriched for microsatellites. *Biotechniques* 27:500–502
- Hammer K, Gladis TH, Laghetti G, Pignone D (2013) *Cichorium intybus* L. as a root vegetable in Italy and remarks on the infraspecific classification of the cultivated race of this species. *International Journal of AgriScience* 3.12:928-937
- Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405
- Kiær LP, Philipp M, Jørgensen RB, Hauser TP (2007) Genealogy, morphology and fitness of spontaneous hybrids between wild and cultivated chicory (*Cichorium intybus*). *Heredity* 99:112–120
- Kiær LP, Felber F, Flavell A, Guadagnuolo R, Guiatti, D, Hauser TP, Olivieri AM, Scotti I, Syed N, Vichi M, van de Wiel C, Jørgensen, RB (2009) Spontaneous gene flow and population structure in wild and cultivated chicory, *Cichorium intybus* L. *Genet Resour Crop Evol* 56:405–419
- Kiers AM (2000) Endive, chicory, and their wild relatives, a systematic and phylogenetic study of *Cichorium* (Asteraceae). PhD thesis, Universiteit Leiden, Leiden
- Kiers AM, Mes THM, van der Meijden R, Bachmann K

- (1999) Morphologically defined *Cichorium* species reflect lineages based on chloroplast and nuclear (ITS) DNA data. *Syst Bot* 24:645–659
- Kiers AM, Mes THM, van der Meijden R, Bachmann K (2000) A search for diagnostic AFLP markers in *Cichorium* species with emphasis on endive and chicory cultivar groups. *Genome* 43:470–476
- Koch G, Jung C (1997) Phylogenetic relationships of industrial chicory varieties revealed by RAPDs and AFLPs. *Agronomie* 17: 323–333
- Lucchin M, Varotto S, Barcaccia G, Parrini P (2008) Chicory and endive: genetics and breeding. In: Prohens J, Carena MJ, Nuez F (eds) *Vegetables I: Asteraceae, Brassicaceae, Chenopodiaceae, and Cucurbitaceae*. Springer, Berlin, pp. 3–48
- Muys C, Thienpont C-N, Dauchot N, Maudoux O, Draye X, Van Cutsem P (2013) Integration of AFLPs, SSRs and SNPs markers into a new genetic map of industrial chicory (*Cichorium intybus* L. var. *sativum*). *Plant Breed* 133:130–137
- Nei M (1973) The theory and estimation of genetic distance. In: Morton NE (ed.) *Genetic Structure of Populations*. University Press of Hawaii, Honolulu, pp. 45–54
- Nei M (1972) Genetic distance between populations. *Amer Nat* 106:283–292
- Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- R Development Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>

- Rick CM (1953) Hybridization between chicory and endive. *Proc Amer Soc Hort Sci* 61:459–466
- Rumball W (1986) Grassland Puna's chicory (*Cichorium intybus* L.). *N Z J Exp Agric* 14:105–107
- Sørensen BS, Kiær LP, Jørgensen RB, Hauser TP (2007) The temporal development in a hybridizing population of wild and cultivated chicory (*Cichorium intybus* L.). *Mol Ecol* 16:3292–3298
- Van Stallen N, Noten V, Neefs V, de Proft M (2001) The phylogenetic relationship between different *Cichorium intybus* cultivars and cultivar groups, as revealed by RAPDs. *Plant Breed* 120:425–428
- Van Cutsem P, du Jardin P, Boutte Beauwens T, Jacqmin S, Vekemans X (2003) Distinction between cultivated and wild chicory gene pools using AFLP markers. *Theor Appl Genet* 107:713–718
- Van Kruistum G (1997) Witlof en roodlof. *Teelthandleiding* 79:11–13.
- Vermeulen A, Desprez B, Lancelin D, Bannerot H (1994) Relationships among *Cichorium* species and related genera as determined by analyses of mt RFLPs. *Theor Appl Genet* 88:159–166

5.9 Supplementary materials

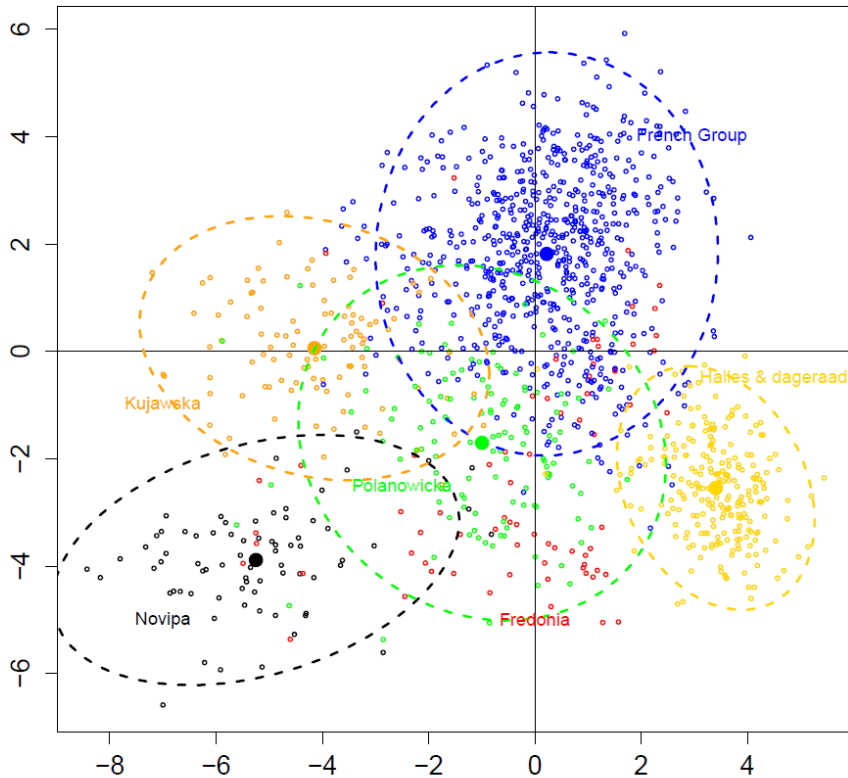
Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the gene pool of industrial chicory



Supplementary material 1: (A & B) Principal component analysis (PCA) of the *C.*

Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the gene pool of industrial chicory

endivia and *C. intybus* cultivar groups (first panel). The red square corresponds to the wild *C. intybus* individuals from Iraq, Pakistan, and the Syrian Arab Republic. (C) PCA of the Witloof group only. (D) PCA of the leaf chicory cultivars.



Supplementary material 2: Principal component analysis of the 15 modern industrial chicory cultivars (second panel). Numbers in ellipses represent the 95% confidence interval of the different groups.

6 Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest inulin depolymerization

This is an adapted version of the research article entitled:

Dauchot, N., Raulier, P., Maudoux, O., Notté, C., Bertin, P., Draye, X., & Van Cutsem, P. (2014). Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest inulin depolymerization. *Theoretical and applied genetics*, 127(1), 125-135.

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest
inulin depolymerization

6.2 Abstract

Industrial chicory (*Cichorium intybus* L.) root is the main commercial source of inulin, a linear fructose polymer used as dietary fiber. Post-harvest, inulin is depolymerized into fructose which drastically increases processing cost. To identify genetic variations associated with enhanced susceptibility to post-harvest inulin depolymerization and related free sugars content increase, we used a candidate-gene approach focused on inulin and sucrose synthesis and degradation genes, all members of the family 32 of glycoside hydrolases (GH32). Polymorphism in these genes was first investigated by carrying out EcoTILLING on two groups of chicory breeding lines exhibiting contrasted response to post-harvest inulin depolymerization. This allowed the identification of polymorphisms significantly associated with depolymerization in three fructan exohydrolase genes (FEH). This association was confirmed on a wider panel of 116 unrelated families in which the FEH polymorphism explained 35 % of the post-harvest variance for inulin content, 36 % of variance for sucrose content, 18 % for inulin degree of polymerization, 23 % for free fructose content and 22 % for free glucose content. These polymorphisms were associated with significant post-harvest changes of inulin content, inulin chain length and free sugars content.

6.3 Introduction

Fructans are the second most abundant reserve carbohydrate among higher plants. They are linear or branched fructan polymers (Vijn and Smeekens 1999) used as an alternative to starch by approximately 15% of flowering plants but also by some bacteria and fungi (Hendry 1993). Inulin, the main commercial fructan polymer, consists of a linear chain of beta 2-1 linked fructosyl units with a single terminal glucose. Inulin is used by the food industry as fat replacer (Arcia et al. 2011; Nowak et al. 2007) or low calories sweetener depending on the length of inulin molecules, while the pharmaceutical industry exploits its health promoting properties (Clark et al. 2012; Closa-Monasterolo et al. 2013; Di Bartolomeo et al. 2013).

Average inulin polymerization degree (DP) varies between species but also according to developmental stage and environmental conditions. In vivo, average inulin polymerization degree (M_{DP}) in *Cynara scolymus* is 8-10 while M_{DP} is around 65 in *Helianthus tuberosus* (Hellwege et al. 1998). In chicory, M_{DP} ranges between 10-20 (Flamm et al. 2001) and is highly influenced by harvest date. Recently, Van Arkel described three phases in inulin metabolism: (I) the onset of inulin biosynthesis one month after sowing with accumulation of high DP inulin until mid-September; (II) from mid-September to early November, gradual decrease of average inulin DP resulting from the synthesis of shorter inulin molecules by back-transfer activity of fructan fructan 1-fructosyltransferase enzyme (1-FFT) (2.4.1.100) and a gradual increase in sucrose concentration; (III)

inulin breakdown following exposure to cold and sub-zero temperatures associated with onset of fructan-exohydrolase activity (FEH), coupled to increase in fructose and sucrose concentrations (van Arkel et al. 2012). Similar findings relative to the variation of inulin chain length during growing season but also post-harvest were reported for *Helianthus tuberosus* (Clausen et al. 2012; Saengthongpinit and Sajjaanantakul 2005) and *Cynara scolymus* (Leroy et al. 2010). Post-harvest fructan depolymerization might be considered as positive in *C. scolymus* because of the reduction of intestinal undesired side effects related to consumption of long fructan chains (Leroy et al. 2010) while this is clearly a problem in industrial chicory (*Cichorium intybus*) which is the main commercial source of inulin. In chicory, Van den Ende reports that in autumn, fructan degrading enzymes (FEH) are activated which increases free fructose and reduces average inulin chain length. High fructose concentration also results in the generation of reducing glucose-free inulo-*n*-ose series which can participate in the Maillard reaction with amino acids during processing. As a result, for large-scale inulin production, early harvest and a quick processing is necessary (Van den Ende and Van Laere 2002). Autumn corresponds to a change in day-length and exposure to cold and sub-zero temperatures. Leaf spoilage by cold is comparable to artificial defoliation performed at harvest, both resulting in a decrease of assimilates from source (leaves) to sink organs (taproot). These parameters were investigated in the context of end-season, post-harvest and storage to understand activation of inulin degrading enzymes. Cold temperatures and frost seem to be responsible for the activation of inulin degrading enzymes, independently of day-length, defoliation (Van den Ende and Van Laere 2002) or water stress (Vandoorne et al. 2012) as confirmed in other species (Leroy et al. 2010; Portes et al. 2008; Saengthongpinit and Sajjaanantakul 2005). In chicory,

Northern blot studies showed that depolymerization is correlated to an increase of 1-FEH I (3.2.1.153) and 1-FEH II (3.2.1.153 and 3.2.1.80) transcripts (Michiels et al. 2004; Van den Ende et al. 2000; Van Laere and Van den Ende 2002) as well as to a decreased transcription of sucrose:sucrose 1-fructosyl transferase (1-SST) (2.4.1.99), the enzyme responsible for the initiation step of fructan biosynthesis. These results were correlated with enzymatic activities, evidencing a transcriptional regulation of inulin metabolism enzymes in chicory (Van den Ende and Van Laere 2002). However, 1-FEH IIa is expected to be more specifically regulated by cold (Van Laere and Van den Ende 2002).

In addition to harvest date, inulin size reduction in chicory following exposure to cold temperature is significantly cultivar-dependent (Baert 1997). In the last years, advances have been achieved in the understanding of the sequence-activity relationship of the members of the family 32 glycoside hydrolases (Van den Ende et al. 2009a; Verhaest et al. 2005), but they do not explain the genetic basis of inter-varietal variability of inulin content and properties commonly exemplified in the official varietal field tests reports (SPW - Portail de l'Agriculture wallonne 2013; van den Brink and Wilting 2012; Vlanderen - Landbouw en Visserij 2013). The first attempt to tackle this question was published recently and concerns the expression levels of 1-FEH IIa/b group which appears to be variable between individuals and tissues of a same chicory cultivar (Maroufi et al. 2010) as well as between cultivars (Maroufi et al. 2012). However, in this last paper, correlation of FEH q-RT PCR data with post-harvest carbohydrates contents and properties were not investigated nor the possible role of other GH32.

In an effort to identify the genetic bases of inter-individual variability on the susceptibility to post-harvest inulin depolymerization, we first looked for nucleotidic polymorphisms within nine fructosyltransferases from the GH32 family: they namely includes sucrose degrading genes (invertases), inulin synthesis (1-SST and 1-FFT) and inulin degrading genes (1-FEH). We then tried to identify any statistical association between the nucleotidic polymorphisms and free sugars and inulin content and properties taking advantage of two phenotypically contrasted breeding lines and later on, for confirmation purpose, of a large collection of 116 unrelated industrial chicory lines. Our results support the implication of the FEH multigenic family.

6.4 Material and methods

6.4.1 Sample sets

Two sample sets were used during this study.

The “first sample set” was used to investigate the nucleotidic polymorphism present within the GH32 by EcoTILLING. This set consists of two groups of respectively eight and twelve breeding lines (Chicoline®) selected for their contrasting susceptibility to post-harvest depolymerization of inulin. One was considered as depolymerization susceptible (DS) while the other was considered as resistant to post-harvest inulin depolymerization (DR). The 20 breeding lines were sown on May 4th, 2007 in Pecq (7740, Belgium) and harvested on November 12th, 2007. Twelve plants per line were harvested following exposure to low temperatures. Ten roots per line were pooled and mixed together for carbohydrates characterization (see dedicated section). The two remaining individuals were sampled for DNA extraction and further EcoTILLING analyses.

A second sample set was used to further investigate the association of the polymorphisms detected within the FEHs with the susceptibility to post-harvest inulin depolymerization. These lines were created after self-pollination of 116 unrelated individuals selected out of a collection of 600 individuals originating from 18 ancient chicory varieties kept in collection by Chicoline®. Selection was realized based on 15 SSR using the MStrat algorithm (Gouesnard et al. 2001).

This sampling algorithm is based on molecular markers to maximize the allelic richness and genetic diversity of a group of sampled individuals. Therefore this strategy minimizes the structure and maximizes the genetic diversity of the selected individuals. Leaf samples of the 116 parents were collected for genomic DNA extraction for the 1-FEH IIa genotyping. The 116 lines were sown on May 2nd, 2011 in Warcoing (7740, Belgium) and 6 to 24 roots per line were harvested between the 2nd and the 9th October, 2011.

6.4.2 Plant Material and DNA extraction

Genomic DNA was extracted from chicory root or leaf powder (manual grinding in liquid nitrogen) according to a typical CTAB 2X procedure (Murray and Thompson 1980) with two additional PCI extractions (Phenol 25: Chloroform 24: Isoamyl alcohol 1).

6.4.3 Carbohydrate characterization

For all the samples used in this study, polysaccharides analyses were performed according to Van Waes (Van Waes et al. 1998). For each breeding line, carbohydrate characterization was performed at harvest and after a cold exposure.

The data were used to determine the following parameters where TF stands for total fructose (after hydrolysis), TG - total glucose (after hydrolysis), FF - free fructose, FG - free glucose and Suc for sucrose (expressed in percentage of fresh

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest inulin depolymerization

weight):

Inulin content (fresh weight) $INc = ((TF + TG - FF - FG - Suc) / 1.11)$
(% of fresh weight)

Carbohydrate content $CC = INc + FF + FG + Suc$ (% of fresh weight)

Polymerization degree $DPin = ((TF - FF - 0.5265 Suc) / (TG - FG - 0.5265 Suc)) + 1$

Inulin content $IN = INc / CC$ (% of Carbohydrate)

Free fructose content $FFn = FF / CC$ (% of Carbohydrate)

Free glucose content $FGn = FG / CC$ (% of Carbohydrate)

Sucrose content $SUCn = Suc / CC$ (% of Carbohydrate)

Dry matter contents were determined starting from grated chicory roots. Each sample was analyzed in duplicate. Chicory root samples were grated and homogenized. 40 gr of grated chicory root were added into pre-weighted evaporating dish and left at 102°C for 24 hours. After 24 hours, samples were cooled down for 2 minutes at room temperature and weighed (evaporating dish and dry material). Samples were set back to the drying cabinet for another 48 hours and weight was compared to the 24h data point to ensure full dehydration.

6.4.4 Primer design and PCR amplification

PCR primers were designed manually from cDNA sequences. Thermodynamic parameters were checked with VNTI software (Invitrogen). Primer pairs were tested on genomic DNA. Optimal annealing temperatures for all amplicons were

determined with a gradient thermal cycler against cloned reference sequences to avoid background or cross amplification between targets. Table 10 lists the primer pairs used for each target region as well as the amplification size on genomic DNA. For automated detection of EcoTILLING digestion products, a single primer of each primer pair was labeled with Wellred D2, D3 or D4 infrared dyes (5' end). PCRs were performed on chicory genomic DNA in 20 µl reaction volumes in 96 wells plates (Biorad iCycler thermocycler) with 50 ng of gDNA, 4 ppm of each primer, 0.6 units of GoTaq DNA Polymerase (Promega) using the provided colorless 5X GoTaq buffer and extension time of 45 seconds. Annealing temperatures and number of cycles are also detailed in Table 10.

Target	Accession number	Size gDNA (bp)	Primer sequence	Ann. T° (°C)	Cycles
1-FEH I	AJ242538	652	F-GTCCACTCTGGGGCAACATTTTC R-D4- TAACCATGAAGTCGTTAGTAACACAG	68.0	40
1-FEH IIa	AY323935	472	F- GTTATGGTGCATTTGTTGACATAGATCC R-D4- ACTTTACTTTATAACAAGAATACTCTAC	59.8	40
1-FEH IIb	AJ295034	591	F- GTTATGGTGCATTTGTTGACATAGATCC R-D4- GTTACAATGGCATCATGGCATTTCATG	59.8	40
1-SST	U81520	606	F-TACGACATCGAAGGGGTCAT R-D2-GACGCATAGAACCTTTCCGTAG	55.0	35
1-FFT	U84398	835	F-CTTATGGTCCCGATATGAAGCA R-D4- ATTCCGAGAGTGTTCCATCAG	55.0	35
Putinv	Y11124	665	F-GGCCCCACTCTGGAATTTACGAATG R-D3- CTAAGCGTTAATTCGTCTTCAGGAAC	61.1	40
CiFr03	JQ082515	629	F-ATCCAAACTCGGTCTCGGCCATTC R-D3-GGGTTAATCTCTTCCGGTAGGTCA	64.0	35
CiFr06	JQ082517	464	F-D3-TCCGTGCTTGGTTGTTTTAG R-GCAATCAATCAGGAGCAATG	60.0	36
CiFr07	JQ082518	655	F-CACATGGGGACATGCAATATCAACG R-D3-AAGTCGTAACCCAATTTCCACATC	61.1	40

Table 10: List of primers and cycling conditions used for EcoTILLING of nine partial sequences of chicory GH32 genes. F and R refer to Forward and Reverse primers and D2, D3 and D4 are the references of the Wellred dyes used for primer labeling before analysis on Beckman Ceq8000 capillary fragment analyzer.

6.4.5 EcoTILLING

EcoTILLING aims to analyze the natural polymorphism present inside genes or genomic regions (Comai et al. 2004). EcoTILLING was performed on selected functional candidate genes: 1-FEH I (AJ242538), 1-FEH IIa (AY323935) and 1-FEH IIb (AJ295034), three fructan exohydrolases, 1-SST (U81520) and 1-FFT (U84398), the two enzymes respectively involved in the initiation and elongation steps of inulin synthesis, Putinv (Y11124), a chicory putative invertase and finally CiFr03 (JQ082515), CiFr06 (JQ082517) and CiFr07 (JQ082518), three chicory GH32 that we recently identified and are believed to be invertases.

For EcoTILLING, we designed an enzyme-based SNP detection assay using a Single Strand Specific nuclease (SSS nuclease). SSS nucleases are sequence-nonspecific enzymes that recognize mispairings in a chimeric DNA molecule obtained after denaturation and annealing of two allelic PCR products. Here, EcoTILLING was performed in two steps. The first step aimed at discriminating homo- from heterozygous samples. The amplification of the target gene was directly followed by denaturation and re-annealing. The labeled amplification products were partially digested with a single strand specific nuclease (Endo-1, Serial Genetics). Labeled digestion fragments were then separated and analyzed on a CEQ8000 capillary sequencer (Beckman Coulter) using MapMarker 1000 as internal size standard (Bioventure, MM-D1 50-1000 bp). In a second step, samples were re-analyzed against a reference PCR product (from a cloned PCR fragment or from an individual homozygous for the analyzed region). Results of the two tests were then merged to generate a deduced co-dominant genotype.

6.4.6 DNA sequencing

Sequencings of FEH alleles were performed on cloned PCR products and outsourced to Macrogen Inc. according to sample submission protocol (Macrogen, The Netherlands).

6.4.7 Genotyping

According to statistical results obtained on the first sample set and subsequent sequencing data, a genotyping primer pair was designed to amplify a 47 bp indel located in the 3'UTR of 1-FEH IIa. This tool was used to genotype 116 unrelated chicory breeding lines to confirm the statistical association between SNPs in FEH sequences and the carbohydrates contents and properties after cold exposure. The target indel was amplified with primers F-GTT-ATG-GTG-CAT-TTG-TTG-ACA-TAG-ATC-C and R-ACT-TTA-CTT-TAT-AAC-AAG-AAT-ACT-CTA-C for 40 cycles at 59.8°C on a Bio-Rad iCycler thermal cycler. Denaturation, annealing and extension were performed for 30 seconds. Duration of initial denaturation was 5 min at 95°C and final extension of 7 minutes at 72°C. PCR reactions were performed in 20 µl volumes and were typically composed of 0.6 units of GoTaq polymerase (Promega), 4 µl 5x GoTaq Buffer (Promega), 4 pm of each primer, 20 pm of each dUTP and 50 ng of gDNA. Amplicons were 6-FAM labeled and were resolved on ABi 3130XL sequencer (Applied Biosystems). Outputs were scored by eye.

6.4.8 Data analysis

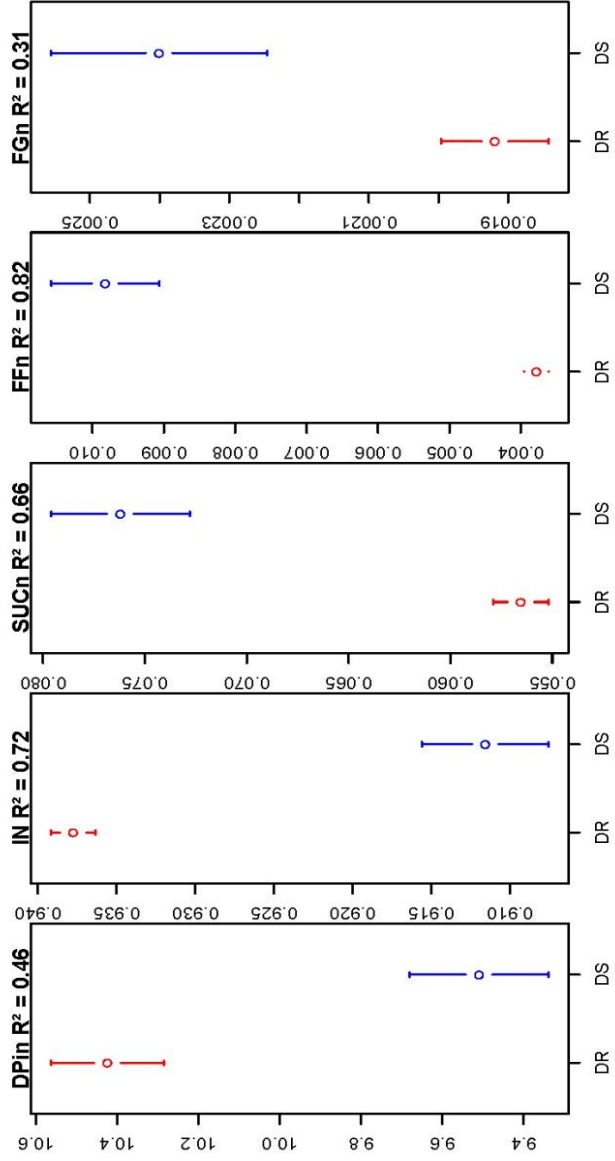
The differences of the different carbohydrates contents and properties (inulin and free sugars) of the two susceptibility groups (first sample set) were analyzed by ANOVA, using the `aov` function of R version 3.0.0. (R Development Core Team 2013). The potential association between the SNP loci identified by EcoTILLING and the differences in carbohydrates contents and properties was indirectly evaluated by comparing the allelic frequencies of the SNP of the two susceptibility groups with a chi square test, using the `chisq.test` function of R version 3.0.0. (R Development Core Team 2013).

The association between the carbohydrates characteristics and the genotyping results obtained on the second sample set with the 1-FEH IIa marker were analyzed by ANOVA, using the `aov` function of R version 3.0.0. (R Development Core Team 2013). Multiple comparisons by means of Tuckey were realized using the function `HSD.test` from the `agricolae` package of R (De Mendiburu 2010).

6.5 Results

6.5.1 Polymorphism and association study

Twenty industrial chicory breeding lines exhibiting contrasted susceptibility to post-harvest inulin depolymerization were used for EcoTILLING of nine members of the GH32 multigenic family. Eight lines were considered as depolymerization susceptible (DS) and twelve were considered as resistant to cold induced inulin depolymerization (DR). These two subsets of breeding lines will be referred to as two “susceptibility groups”. A characterization of the carbohydrate contents and properties was first performed on these two phenotypic groups to ensure phenotypic variability between the two groups. Results confirmed that, compared to the DR lines, the DS lines presented, post-harvest, significantly different carbohydrate contents with pValues all lower than 0.006. The classification of the 20 breeding lines in two susceptibility groups (DS and DR) explained a high proportion of the variance of the different parameters (DPin, IN, SUCn, FGn and FFn) with R^2 ranging from 0.31 to 0.82 (Figure 38).



Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest
inulin depolymerization

Figure 38: Means and standard error of eight depolymerization susceptible lines and the twelve depolymerization resistant lines (DS vs DR , first sample set) for the post-harvest degree of polymerization of inulin (DPin), the inulin content (IN), the sucrose content (SUCn), the free fructose content (FFn) and the free glucose content (FGn). All phenotypic parameters are significantly different between the susceptibility two groups ($P < 0.005$).

EcoTILLING identified 22 SNPs and 4 insertions/deletions (Table 11). These results were compared to carbohydrate related phenotypic groups (DS vs DR lines) to look for statistical association (Table 12). The loci 1-FEH I 76, 97, 153, 1-FEH IIa 142, 190, 301, 1-FEH IIa amplification product (430/473) and 1-FEH IIb 94 showed a clear cut segregation of the SNP alleles between the DS and DR breeding lines: SNP alleles were strictly identical within all susceptible line (DS), but were different from those of the resistant lines (DR) with χ^2 p-value of 3×10^{-18} . The two last polymorphic loci located within the 1-FEH IIb gene (178, 234) were also significantly correlated but with slightly different p-values for the χ^2 test ranging from 1.3×10^{-9} to 5.2×10^{-10} . None of the others polymorphic loci analyzed by EcoTILLING showed any significant correlation with the two susceptibility groups (DS/DR). Fructan exohydrolases are therefore good candidates for inter-individual variability to cold-induced inulin depolymerization.

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest inulin depolymerization

Table 11: Results of EcoTILLING analyzes on 9 chicory GH32s performed on 20 chicory breeding lines exhibiting contrasting susceptibility to post-harvest inulin depolymerization. Two individuals per breeding line were analyzed, respectively A and B at the end of the name of the sample (first column). The samples were depolymerization susceptible (DS) or depolymerization resistant (DR) in the second column. Each column corresponds either to (i) a SNP position detected after scoring of the EcoTILLING digestion profiles or (ii) to the estimated size of the undigested amplification products as listed in the “AP” columns. SNP results are presented as « AA » or « BB » when homozygous or « AB » when heterozygous.

TARGET	LOCUS	χ value
FEHI	76	75.9 ***
FEHI	97	75.9***
FEHI	153	75.9***
FEHI	472	38.6***
FEHIIa	142	75.9***
FEHIIa	190	75.9***
FEHIIa	301	75.9***
FEHIIaAP	430/473	75.9***
FEHIIa	94	75.9***
FEHIIa	178	36.7***
FEHIIa	234	36.7***
1-SST	267	3.2
1-FFT	178	3.2
FFT_AP	826/834	3.2
Putinv	100	0.0
Putinv	102	0.0
Putinv	320	0.0
Putinv	585	0.0
CiFr03	223	0.0
CiFr03	406	0.2
CiFr03	600/630	4.2
Cifr06	136	1.4
Cifr06	204	0.0
Cifr06	325	0.4
CiFr06	457/460	0.4
CiFr07	279	4.6

Table 12: Referring to the EcoTILLING results (Table 11). χ^2 values evaluating the distribution of the alleles (26 polymorphic loci) among the two susceptibility group (first sample set: DS vs DR). Statistically significant effect are indicated * $P < 0$

In an attempt to identify a causative mutation among the detected SNPs, the EcoTILLING fragments of the three FEHs with statistically significant SNPs were sequenced on three samples originating from the DS and DR breeding lines. No polymorphism was detected within the susceptibility groups, but several SNPs and indels were identified between the DS and DR lines. Alignments were performed comparing chicory cDNA sequences already present at NCBI and the partial gDNA sequences of 1-FEH I, 1-FEH IIa and 1-FEH IIb obtained during this study. Sequence alignments are presented in (Supplementary material 3, and Supplementary material 5). SNPs were detected in both introns and exons. The published cDNA sequences of the three genes (AJ242538, AJ295033 and AJ295034) almost perfectly matched the alleles identified in DS lines, while they presented several SNP when compared to the alleles present in DR lines.

The 1-FEH I EcoTILLING fragment covers three conserved regions, namely WINDPNG, FRDP and WECPD that contain the three catalytic residues of GH32 enzymes. None of the identified polymorphisms in 1-FEH I was located in these conserved regions (Supplementary material 3). However, four non-silent SNPs were located close to the WECPD sequence which could affect the environment of the E residue in the active site. Attempts to identify, by homology modeling, any putative structural modifications induced by these mutations were unsuccessful (not shown). These two partial genomic sequences are accessible at NCBI under accession numbers JQ585636 (1-FEH-I-R) and JQ585637 (1-FEH-I-S).

1-FEH IIa and 1-FEH IIb EcoTILLING fragments were located across the last

two exons, including a significant part of the 3'UTR region and one short intron.

Out of the six SNP mutations identified in 1-FEH IIa, only one induced a modification of the amino-acid sequence. However, a 47 bp duplication was also observed in the 3'UTR of depolymerization susceptible 1-FEH IIa alleles (D+) compared to depolymerization resistant 1-FEH IIa alleles (d-) (Supplementary material 5). The “long” 1-FEH IIa allele will be referred to as the “D+” allele (presence of the duplication) while the “short” 1-FEH IIa allele will be the “d-” allele (for absence of duplication). This 47 bp region is a perfect tandem repeat located in the long 1-FEH IIa allele (D+ / JQ585639) while a single copy is present in the shorter allele (d- / JQ585638). This 47 bp sequence comprises the six last codons of the cds (including the terminal TAA stop codon) and the 29 first bp of the 3'UTR. The sequence we identified in susceptible breeding lines (JQ585639) was 100% identical to the 3'UTR of previously published AJ295033 while the partial 3' UTR of the alleles identified in depolymerization-resistant lines (JQ585638) contained 2 SNPs within their single copy of the 47 bp region.

Out of the 10 SNP mutations identified in 1-FEH IIb, only two were non-silent while seven were silent and a last one was an indel of 1 bp in an intron (Supplementary material 6). 1-FEH IIb sequences have accession numbers JQ585640 (1-FEH-IIb-S) and JQ585641 (1-FEH-IIb-R).

6.5.2 Confirmation of the association

In absence of any detected causative mutation in the three FEHs examined and considering the likely physical linkage between these three genes, the ease of scoring a 47 bp indel and the previously reported implication of 1-FEH IIa/b in cold-induced inulin depolymerization (Van Laere and Van den Ende 2002), a larger set of 116 unrelated lines was studied by genotyping only the 47 bp indel present in the 3'UTR of 1-FEH IIa to confirm the preliminary detected associations.

Carbohydrate related phenotypic parameters were then evaluated on the progeny of 116 lines, before and after post-harvest cold exposure and compared to 1-FEH IIa genotyping results. At harvest, the average dry mater and carbohydrates content were 24.6% and 17.8% of total fresh weight of both groups of lines. The carbohydrates mainly consisted in inulin (IN = 94.6%) and to a lower proportion of sucrose (SUCn = 4.9%). The free fructose (FFn) and free glucose (FGn) represented only 0.33 and 0.17% of the carbohydrates. The average degree of polymerization (DPin) was 12.2. At this stage, the nature of the 1-FEH IIa genotype was not significantly correlated with any of these parameters ($\alpha = 0.01$) (**Supplementary material 4**).

The dry mater and carbohydrate contents were not significantly different after the cold exposure but we observed (i) a reduced DPin and inulin content (IN), (ii) an important increase in sucrose (SUCn) and fructose (FFn) concentrations and (iii) a slight increase in glucose (FGn) concentration (Figure 39). These

variations were all significantly correlated to the nature of the 1-FEH IIa genotype (p-value ranging from $4.2 \cdot 10^{-6}$ to $4.1 \cdot 10^{-12}$). More precisely, the homozygous (D+/D+) had lost 25.8% of IN and 4 units of DP_{in} after the cold exposure while the homozygous (d-/d-) had lost only 10.8% of IN and 0.69 unit of DP_{in}. The sucrose (SUC_n), fructose (FF_n) and glucose (FG_n) increased by 17.8%, 7.7% and 0.36% respectively for the (D+/D+) genotype, and by only 7.43%, 3.22% and 0.129% for the (d-/d-) genotype. The heterozygous (D+/d-) genotype presented an intermediate phenotype for IN, SS and SUC_n but they clustered with (i) the homozygous (D+/D+) genotype for FF_n and FG_n and (ii) the (d-/d-) genotype for DP_{in}. Presence or absence of the 47 bp indel in 1-FEH IIa explained a high proportion of IN and SUC_n variation following cold storage ($R^2_{IN} = 0.35$ and $R^2_{SUCn} = 0.36$) and a lower proportion for the DP_{in}, FF_n and FG_n ($R^2_{DPin} = 0.18$, $R^2_{FFn} = 0.23$ and $R^2_{FGn} = 0.22$).

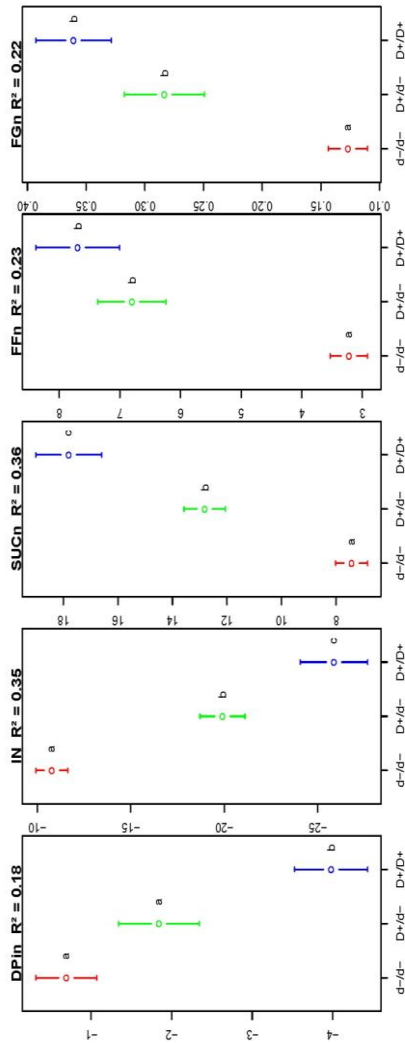


Figure 39: Means and standard error of the three different 1-FEH IIa genotypes (“d-” refers to the allele “absence of duplication” and the “D+” to the allele “presence of the duplication”) among 116 lines for the differences before and after cold exposure

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest inulin depolymerization

of the inulin degree of polymerization (DP_{in}), inulin content (IN), sucrose content (SUC_n), free fructose content (FF_n) and free glucose content (FG_n). The means followed by different letters were significantly different ($P < 0.01$). The 1-FEH IIa genotype significantly affected all five investigated phenotypic parameters ($P < 0.00001$).

6.6 Discussion

Transcriptional regulation of inulin synthesis and degradation genes belonging to the GH32 family of hydrolases in response to late season exposure to cold temperature has been reported (Van den Ende et al. 2002). Even if these results pinpoint clear candidate genes for cold-induced inulin depolymerization, they do not explain inter-individual variability of susceptibility to post-harvest inulin depolymerization.

In the past, the importance of point mutations on the activity of GH32 was investigated and residues implicated in the active site, substrate specificity and regulation were described (Ritsema et al. 2006; Van den Ende and Valluru 2009; Verhaest et al. 2007). However, the relationship between GH32 point mutations and the susceptibility to post-harvest inulin depolymerization was not identified by these authors.

This research was composed of two distinct, but complementary, steps. A first step consisted into the investigation of the natural polymorphism present among several representative members of the GH32 multigenic family in chicory. We then evaluated the possible association between the identified polymorphisms and post-harvest phenotypic variability of chicory root carbohydrates content and properties. This first step was conducted on a limited number (20) of phenotypically contrasted breeding lines. This first step allowed identifying statistically significant polymorphisms within three strong candidate genes, namely 1-FEH I, 1-FEH IIa and 1-FEH IIb. In the second part of this research,

based on these preliminary results, a genotyping probe targeting an indel located inside 1-FEH IIa was developed and used on a collection of 116 unrelated chicory lines selected to maximize genetic diversity and minimize genetic structure to confirm the detected genotype/phenotype associations.

Inulin chain length in the root is believed to be affected by both synthesis and depolymerization mechanisms (Van den Ende et al. 1996). However, our results did not identify any statistical association between 1-SST or 1-FFT polymorphisms, the two key enzymes responsible for the initiation and elongation of inulin, and the post-harvest average inulin chain length or free sugars content.

At the level of degrading enzymes, two distinct classes were investigated, namely invertases and fructan exohydrolases. Invertases are sucrose cleaving enzymes. Their activity modifies sucrose availability which could affect synthesis process, and increase free glucose and free fructose. Fructan exohydrolases are inulin degrading enzymes. Their activity results into a general decrease of the average inulin chain length and into an increase of the free fructose and sucrose.

Based on its high sequence homology to a *Daucus carota* beta-fructofuranosidase (X75352/CAA53098), CiFr07, a chicory GH32 that we recently identified, could be an acid vacuolar invertase. The unique SNP identified in CiFr07 was not associated with a post-harvest difference of the degree of polymerization of inulin (DP_{in}) or small sugar content (FG_n, FF_n, SUC_n).

The last six sequences that were investigated all belong to the group of FEH/cell wall invertases. Fructan hydrolysis genes are believed to have evolved from cell wall invertase and gained sucrose inhibition (1-FEH IIa and IIb) with specificity for β -2,1 linear fructan molecules. 1-FEH I is also a fructan hydrolyzing enzyme, but with a weaker susceptibility to sucrose inhibition. For this study, we also included two new chicory GH32, CiFr03 and CiFr06 that are likely to be cell wall invertases since they exhibit the S101 to I101 substitution in the GSAT - YTG conserved regions, which is associated with the absence of sucrose inhibition (Verhaest et al. 2007). The presence the D - - K/R motif specific to cell wall invertases also supports this assignation (Van den Ende et al. 2009b). Polymorphisms detected in CiFr03 and CiFr06 lacked statistical association to phenotypic data. Putinv (Y11124) was the sequence presenting the highest homology to CiFr03 and CiFr06 (80%) and did not show any statistical association either.

We found strong statistical association in the first small set of genotypes between nucleotidic polymorphisms located within three FEHs (1-FEH I, 1-FEH IIa and 1-FEH IIb) and carbohydrate parameters. Partial genomic sequencing of these three FEHs identified a 47 bp indel within the 3'UTR of 1-FEH IIa. Based on *in silico* secondary structure predictions, one can speculate that the presence of a perfect tandem repeat in the 3' UTR of 1-FEH IIa could generate a second hairpin structure (Supplementary material 7) and possibly affect m-RNA stability or pre-mRNA processing as reported previously (Gutierrez et al. 1999; Millevoi and Vagner 2009).

We also observed an apparent linkage of the genotyping results observed between the three FEHs (Table 11) suggesting a possible physical proximity of the three genes. Previous results reported physical proximity of 1-FEH IIa and 1-FEH IIb located at 28.8 cM and 30.6 cM respectively on LG4 of the Rubis 118 physical map of chicory (Cadalen et al. 2010). In lettuce (CGPDB 2013), 1-FEH I and 1-FEH II genes are located about 40 kb from each other (CGBPD, Davis. 1-FEH I, LG5 116059887-116060805, 1-FEH II LG5 116092584-116093466), which, considering the existing synteny between the two species (Muys et al. 2013), might support the physical proximity of the three FEHs. Considering the high redundancy observed between statistically significant SNP loci located in the three FEHs, taking into account previous data supporting the role of 1-FEH II in cold-induced inulin depolymerization and considering the ease of scoring a large indel, we decided to focus on 1-FEH IIa.

The statistical association between 1-FEH II and post-harvest inulin depolymerization identified on the first sample set was confirmed on a second, wider, genetically diversified and unstructured collection of 116 industrial chicories lines. In this study, total carbohydrate content was identical before and after cold storage while inulin and free sugars contents were modified. Post-harvest, the 1-FEH IIa marker explained approximately 35% of the variance of inulin and sucrose contents. For these carbohydrates, the three genotypes (d-/d-, D+/d- and D+/D+) were statistically different, suggesting a co-dominant effect of the causative loci. This additive effect was less clear for the inulin degree of polymerization (DPin), free fructose (FFn) and free glucose (FGn) contents for which R^2 values were around 0.2.

An intriguing result of this study concerns the free carbohydrates content post-harvest. The drop in inulin content resulted in an equivalent increase of free sugars. However, free fructose and to a smaller extent sucrose (end product of inulin depolymerization by FEH enzymes) should have increased much more than observed. Fructose increase represented only 1/3 of the increase, while sucrose accounted for 2/3 of the post-harvest increase in free sugars, coupled with a very low level of free glucose (between 0.1 and 0.4%). This raises a question about recycling of the free fructose generated during inulin hydrolysis. Free fructose might have been converted to sucrose by sucrose synthase (EC 2.4.1.13) or sucrose phosphate synthase (EC 2.4.1.14), resulting in low levels of free glucose and an increase in sucrose (Nguyen-Quoc and Foyer 2001).

6.7 Conclusions

A candidate-gene approach was successfully applied to identify phenotype/genotype association in the non-model species chicory. We identified several SNPs and indel polymorphisms located in three FEH genes and found a statistical association between a 47 bp indel located in the 3'UTR of 1-FEH IIa and an enhanced susceptibility to post-harvest inulin depolymerization. These results are of particular interest if we consider that these FEHs are, until now, the only functionally characterized β -2,1- hydrolyzing enzymes identified in chicory. Statistical analyses indicated that identified polymorphisms could explain up to two thirds of the inulin loss and free sugars accumulation and save up to 5 DP units on the average inulin chain length post-harvest.

Acknowledgements

The authors acknowledge the Walloon Region (DGARNE-Belgium) for supporting this research (grant D31-1221).

Conflict of interest

O. M. and C. N. are members of Cosucra – Group Warcoing S.A.

Ethical standards

The authors acknowledge that the experiments described in this paper comply with the current laws of the country in which they were performed

6.8 Reference

Arcia PL, Costell E, Tarrega A (2011) Inulin blend as prebiotic and fat replacer in dairy desserts: optimization by response surface methodology. *J Dairy Sci* 94:2192-2200

Baert JRA (1997) The effect of sowing and harvest date and cultivar on inulin yield and composition of chicory (*cichorium intybus* L.) roots. *Industrial Crops and Products* 6:195-199

Cadalen T, Morchen M, Blassiau C, Clabaut A, Scheer I, Hilbert JL, Hendriks T, Quillet MC (2010) Development of SSR markers and construction of a consensus genetic map for chicory (*Cichorium intybus* L.). *Molecular Breeding* 25:699-722

CGPDB (2013) Lettuce Genome Resource

<https://lgr.genomecenter.ucdavis.edu/Home.php>. Accessed online 2013/05/20

Clark MJ, Robien K, Slavin JL (2012) Effect of prebiotics on biomarkers of colorectal cancer in humans: a systematic review. *Nutr Rev* 70:436-443

Clausen MR, Bach V, Edelenbos M, Bertram HC (2012) Metabolomics reveals drastic compositional changes during overwintering of Jerusalem artichoke (*Helianthus tuberosus* L.) tubers. *J Agric Food Chem* 60:9495-9501

Closa-Monasterolo R, Gispert-Llaurado M, Luque V, Ferre N, Rubio-Torrents C, Zaragoza-Jordana M, Escribano J (2013) Safety and efficacy of inulin and oligofructose supplementation in infant formula: Results from a randomized clinical trial. *Clin Nutr*

Comai L, Young K, Till BJ, Reynolds SH, Greene EA, Codomo CA, Enns LC, Johnson JE, Burtner C, Odden AR, Henikoff S (2004) Efficient discovery of

DNA polymorphisms in natural populations by Ecotilling. *Plant Journal* 37:778-786

De Mendiburu F (2010) *Agricolae: Statistical Procedures for Agricultural Research Using R* (Online)

<http://tarwi.lamolina.edu.pe/~fmendiburu/>. Accessed online 2013/03/05

Di Bartolomeo F, Startek JB, Van den Ende W (2013) Prebiotics to Fight Diseases: Reality or Fiction? *Phytother Res*

Flamm G, Glinsmann W, Kritchevsky D, Prosky L, Roberfroid M (2001) Inulin and oligofructose as dietary fiber: a review of the evidence. *Crit Rev Food Sci Nutr* 41:353-362

Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. *J Hered* 92:93-94

Gutierrez RA, MacIntosh GC, Green PJ (1999) Current perspectives on mRNA stability in plants: multiple levels and mechanisms of control. *Trends Plant Sci* 4:429-438

Hellwege EM, Raap M, Gritscher D, Willmitzer L, Heyer AG (1998) Differences in chain length distribution of inulin from *Cynara scolymus* and *Helianthus tuberosus* are reflected in a transient plant expression system using the respective 1-FFT cDNAs. *FEBS Lett* 427:25-28

Hendry GAF (1993) Evolutionary Origins and Natural Functions of Fructans - a Climatological, Biogeographic and Mechanistic Appraisal. *New Phytologist* 123:3-14

Leroy G, Grongnet JF, Mabeau S, Corre DL, Baty-Julien C (2010) Changes in inulin and soluble sugar concentration in artichokes (*Cynara scolymus* L.) during storage. *J Sci Food Agric* 90:1203-1209

- Maroufi A, Van Bockstaele E, De Loose M (2010) Validation of reference genes for gene expression analysis in chicory (*Cichorium intybus*) using quantitative real-time PCR. *BMC Mol Biol* 11:15
- Maroufi A, Van Bockstaele E, De Loose M (2012) Differential expression of fructan 1-exohydrolase genes involved in inulin biodegradation in chicory (*Cichorium intybus*) cultivars. *Australian Journal of Crop Science* 6:1362-1368
- Michiels A, Van Laere A, Van den Ende W, Tucker M (2004) Expression analysis of a chicory fructan 1-exohydrolase gene reveals complex regulation by cold. *Journal of Experimental Botany* 55:1325-1333
- Millevoi S, Vagner S (2009) Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Res* 38:2757-2774
- Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res* 8:4321-4325
- Muys C, Thienpont C-N, Dauchot N, Maudoux O, Draye X, Van Cutsem P (2013) Integration of AFLPs, SSRs and SNPs markers into a new genetic map of industrial chicory (*Cichorium intybus* L. va. *sativum*). *Plant Breeding* (in press)
- Nguyen-Quoc B, Foyer CH (2001) A role for 'futile cycles' involving invertase and sucrose synthase in sucrose metabolism of tomato fruit. *J Exp Bot* 52:881-889
- Nowak B, von Mueffling T, Grotheer J, Klein G, Watkinson BM (2007) Energy content, sensory properties, and microbiological shelf life of German bologna-type sausages produced with citrate or phosphate and with inulin as fat replacer. *J Food Sci* 72:S629-638
- Portes MT, Figueiredo-Ribeiro Rde C, de Carvalho MA (2008) Low temperature and defoliation affect fructan-metabolizing enzymes in different regions of the rhizophores of *Vernonia herbacea*. *J Plant Physiol* 165:1572-1581

- R Development Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
- Ritsema T, Hernandez L, Verhaar A, Altenbach D, Boller T, Wiemken A, Smeekens S (2006) Developing fructan-synthesizing capability in a plant invertase via mutations in the sucrose-binding box. *Plant J* 48:228-237
- Saengthongpinit W, Sajjaanantakul T (2005) Influence of harvest time and storage temperature on characteristics of inulin from Jerusalem artichoke (*Helianthus tuberosus* L.) tubers. *Postharvest Biology and Technology* 37:93-100
- SPW - Portail de l'Agriculture wallonne (2013) Catalogues nationaux des variétés d'espèces de plantes agricoles et des variétés d'espèces de légumes http://agriculture.wallonie.be/apps/spip_wolwin/article.php3?id_article=68. Accessed online 2013/05/31
- van Arkel J, Vergauwen R, Sevenier R, Hakkert JC, van Laere A, Bouwmeester HJ, Koops AJ, van der Meer IM (2012) Sink filling, inulin metabolizing enzymes and carbohydrate status in field grown chicory (*Cichorium intybus* L.). *J Plant Physiol* 169:1520-1529
- van den Brink L, Wilting P (2012) Rassenbulletin cichorei <http://www.kennisakker.nl/kenniscentrum/document/rassenbulletin-cichorei>. Accessed online 2013/05/31
- Van den Ende W, Lammens W, Van Laere A, Schroeven L, Le Roy K (2009a) Donor and acceptor substrate selectivity among plant glycoside hydrolase family 32 enzymes. *Febs Journal* 276:5788-5798
- Van den Ende W, Lammens W, Van Laere A, Schroeven L, Le Roy K (2009b) Donor and acceptor substrate selectivity among plant glycoside hydrolase family 32 enzymes. *FEBS J* 276:5788-5798
- Van den Ende W, Michiels A, De Roover J, Van Laere A (2002) Fructan

biosynthetic and breakdown enzymes in dicots evolved from different invertases. Expression of fructan genes throughout chicory development. *ScientificWorldJournal* 2:1281-1295

Van den Ende W, Michiels A, Van Wouterghem D, Vergauwen R, Van Laere A (2000) Cloning, developmental, and tissue-specific expression of sucrose:sucrose 1-fructosyl transferase from *Taraxacum officinale*. Fructan localization in roots. *Plant Physiol* 123:71-80

Van den Ende W, Mintiens A, Speleers H, Onuoha A, Van Laere A (1996) The metabolism of fructans in roots of *Cichorium intybus* during growth, storage and forcing. *New Phytol* 132:555-563

Van den Ende W, Valluru R (2009) Sucrose, sucrosyl oligosaccharides, and oxidative stress: scavenging and salvaging? *J Exp Bot* 60:9-18

Van den Ende W, Van Laere A (2002) Induction of 1-FEH in mature chicory roots appears to be related to low temperatures rather than to leaf damage. *ScientificWorldJournal* 2:1750-1761

Van Laere A, Van den Ende W (2002) Inulin metabolism in dicots: chicory as a model system. *Plant Cell and Environment* 25:803-813

Vandoorne B, Mathieu AS, Van den Ende W, Vergauwen R, Perilleux C, Javaux M, Lutts S (2012) Water stress drastically reduces root growth and inulin yield in *Cichorium intybus* (var. *sativum*) independently of photosynthesis. *J Exp Bot* 63:4359-4373

Verhaest M, Lammens W, Le Roy K, De Ranter CJ, Van Laere A, Rabijns A, Van den Ende W (2007) Insights into the fine architecture of the active site of chicory fructan 1-exohydrolase: 1-kestose as substrate vs sucrose as inhibitor. *New Phytol* 174:90-100

Verhaest M, Van den Ende W, Le Roy K, De Ranter CJ, Van Laere A, Rabijns A

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest
inulin depolymerization

(2005) X-ray diffraction structure of a plant glycosyl hydrolase family 32 protein: fructan 1-exohydrolase IIa of *Cichorium intybus*. *Plant Journal* 41:400-411

Vijn I, Smeeckens S (1999) Fructan: more than a reserve carbohydrate? *Plant Physiol* 120:351-360

Vlaanderen - Landbouw en Visserij (2013) Industriële cichorei: resultaten van rassenproeven

<http://lv.vlaanderen.be/nlapps/docs/default.asp?fid=153>. Accessed online 2013/05/31

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest
inulin depolymerization

6.9 Supplementary materials

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest
inulin depolymerization

			181		W I N D P N G	240
1-FEH I - AJ242538	(181)	TACAGGACTGGCTTCCACTTCCAGCCTCCTAAAAAC		TGGATAAACGATCCTAACGGATCC		
1-FEH I (DS) - JQ585637	(1)	-----		-----		
1-FEH I (DR) - JQ585636	(1)	-----		-----		
Consensus	(181)					
			241			300
1-FEH I - AJ242538	(241)	ATGTACTTCAATGGAGTTTACCATCTGTTTACCAATACAACCCCTTACG		GTCCACTCTGG		
1-FEH I (DS) - JQ585637	(1)	-----		GTCCACTCTGG		
1-FEH I (DR) - JQ585636	(1)	-----		GTCCACTCTGG		
Consensus	(241)			GTCCACTCTGG		
			301			360
1-FEH I - AJ242538	(301)	GGCAACATTTTCATGGGGTCATTCCATTTTCATATGATCTAGTGAACCTGGTTTCTTCTCGAA				
1-FEH I (DS) - JQ585637	(12)	GGCAACATTTTCATGGGGTCATTCCATTTTCATATGATCTAGTGAACCTGGTTTCTTCTCGAA				
1-FEH I (DR) - JQ585636	(12)	GGCAACATTTTCATGGGGTCATTCCATTTTCATATGATCTAGTGAACCTGGTTTCTTCTCGAA				
Consensus	(301)	GGCAACATTTTCATGGGGTCATTCCATTTTCATATGATCTAGTGAACCTGGTTTCTTCTCGAA				
			361			420
1-FEH I - AJ242538	(361)	CCAGCTCTCAGTCCAAAAGAACCCTTATGACATCAATGGCTGCCTGTCAGGTTTCAGCCACG				
1-FEH I (DS) - JQ585637	(72)	CCAGCTCTCAGTCCAAAAGAACCCTTATGACATCAATGGCTGCCTGTCAGGTTTCAGCCACG				
1-FEH I (DR) - JQ585636	(72)	CCAGCTCTCAGTCCAAAAGAACCCTTATGACATCAATGGCTGCCTGTCAGGTTTCAGCCACG				
Consensus	(361)	CCAGCTCTCAGTCCAAAAGAACCCTTATGACATCAATGGCTGCCTGTCAGGTTTCAGCCACG				
			421			480
1-FEH I - AJ242538	(421)	ATCCTACCCGGGCCACGGCCATAATTTCTATATACCGGTCAAGACGTAACAATAGTCAG				
1-FEH I (DS) - JQ585637	(132)	ATCCTACCCGGGCCACGGCCATAATTTCTATATACCGGTCAAGACGTAACAATAGTCAG				
1-FEH I (DR) - JQ585636	(132)	ATCCTACCCGGGCCACGGCCATAATTTCTATATACCGGTCAAGACGTAACAATAGTCAG				
Consensus	(421)	ATCCTACCCGGGCCACGGCCATAATTTCTATATACCGGTCAAGACGTAACAATAGTCAG				
			481			540
1-FEH I - AJ242538	(481)	GTTCAGAACCTGGCCTTCCCTAAAAATCTCTCTGACCCCTCTTTTGAAAGAGTGGATAAAA				
1-FEH I (DS) - JQ585637	(192)	GTTCAGAACCTGGCCTTCCCTAAAAATCTCTCTGACCCCTCTTTTGAAAGAGTGGATAAAA				
1-FEH I (DR) - JQ585636	(192)	GTTCAGAACCTGGCCTTCCCTAAAAATCTCTCTGACCCCTCTTTTGAAAGAGTGGATAAAA				
Consensus	(481)	GTTCAGAACCTGGCCTTCCCTAAAAATCTCTCTGACCCCTCTTTTGAAAGAGTGGATAAAA				
			541		F R D	
1-FEH I - AJ242538	(541)	TGGTCGGGTAACCCCTTTGCTGACCCCGCTCGATGACATTAAAGCTGGACAAATTCGGTGAT				
1-FEH I (DS) - JQ585637	(252)	TGGTCGGGTAACCCCTTTGCTGACCCCGCTCGATGACATTAAAGCTGGACAAATTCGGTGAT				
1-FEH I (DR) - JQ585636	(252)	TGGTCGGGTAACCCCTTTGCTGACCCCGCTCGATGACATTAAAGCTGGACAAATTCGGTGAT				
Consensus	(541)	TGGTCGGGTAACCCCTTTGCTGACCCCGCTCGATGACATTAAAGCTGGACAAATTCGGTGAT				
			601		P	660
1-FEH I - AJ242538	(601)	CCATCAACGGCTTGGATGGGTCCGGATGGAAAAATGGAGGATTGTGATCGGAAGTGAGATA				
1-FEH I (DS) - JQ585637	(312)	CCATCAACGGCTTGGATGGGTCCGGATGGAAAAATGGAGGATTGTGATCGGAAGTGAGATA				
1-FEH I (DR) - JQ585636	(312)	CCATCAACGGCTTGGATGGGTCCGGATGGAAAAATGGAGGATTGTGATCGGAAGTGAGATA				
Consensus	(601)	CCATCAACGGCTTGGATGGGTCCGGATGGAAAAATGGAGGATTGTGATCGGAAGTGAGATA				
			661			720
1-FEH I - AJ242538	(661)	GATGGTCATGGAACCGCACTTTTGTATGTAGTAGCAATGGTACAAAGTGAAATCGGATCC				
1-FEH I (DS) - JQ585637	(372)	GATGGTCATGGAACCGCACTTTTGTATGTAGTAGCAATGGTACAAAGTGAAATCGGATCC				
1-FEH I (DR) - JQ585636	(372)	GATGGTCATGGAACCGCACTTTTGTATGTAGTAGCAATGGTACAAAGTGAAATCGGATCC				
Consensus	(661)	GATGGTCATGGAACCGCACTTTTGTATGTAGTAGCAATGGTACAAAGTGAAATCGGATCC				
			721		W E C P D	780
1-FEH I - AJ242538	(721)	AAAAAACCTTTGCACCTTTTCTAGTAAACCGGGTATGTGGAGGTGTCCTGAC				
1-FEH I (DS) - JQ585637	(432)	AAAAAACCTTTGCACCTTTTCTAGTAAACCGGGTATGTGGAGGTGTCCTGAC				
1-FEH I (DR) - JQ585636	(432)	AAAAAACCTTTGCACCTTTTCTAGTAAACCGGGTATGTGGAGGTGTCCTGAC				
Consensus	(721)	AAAAAACCTTTGCACCTTTTCTAGTAAACCGGGTATGTGGAGGTGTCCTGAC				
			781			840
1-FEH I - AJ242538	(781)	GTTACTAATGGTGACAAAAAGGGTCTTGACACATCTGTTCAAGGGAATAACACACTTCAT				
1-FEH I (DS) - JQ585637	(492)	GTTACTAATGGTGACAAAAAGGGTCTTGACACATCTGTTCAAGGGAATAACACACTTCAT				
1-FEH I (DR) - JQ585636	(492)	GTTACTAATGGTGACAAAAAGGGTCTTGACACATCTGTTCAAGGGAATAACACACTTCAT				
Consensus	(781)	GTTACTAATGGTGACAAAAAGGGTCTTGACACATCTGTTCAAGGGAATAACACACTTCAT				

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest inulin depolymerization

97 * 76 * *

		(841)	GTTCTCAAAGTAAGCTTCAATAGTCGTGAGTATTATGTAATCGGCACCTTATGATCCAATA	(900)
1-FEH I - AJ242538	(841)		GTTCTCAAAGTAAGCTTCAATAGTCGTGAGTATTATGTAATCGGCACCTTATGATCCAATA	
1-FEH I (DS) - JQ585637	(552)		GTTCTCAAAGTAAGCTTCAATAGTCGTGAGTATTATGTAATCGGCACCTTATGATCCAATA	
1-FEH I (DR) - JQ585636	(552)		GTTCTCAAAGTAAGCTTCAATAGTCGTGAGTATTATGTAATCGGCACCTTATGATCCAATA	
Consensus	(841)		GTTCTCAAAGTAAGCTTCAATAGTCGTGAGTATTATGTAATCGGCACCTTATGATCCAATA	(900)

		(901)	AAGAAGAACTTTCTTGTTGGTTACTAACGACTTCATGGTTA	(960)
1-FEH I - AJ242538	(901)		AAGAAGAACTTTCTTGTTGGTTACTAACGACTTCATGGTTA	
1-FEH I (DS) - JQ585637	(612)		AAGAAGAACTTTCTTGTTGGTTACTAACGACTTCATGGTTA-----	
1-FEH I (DR) - JQ585636	(612)		AAGAAGAACTTTCTTGTTGGTTACTAACGACTTCATGGTTA-----	
Consensus	(901)		AAGAAGAACTTTCTTGTTGGTTACTAACGACTTCATGGTTA	

Supplementary material 3: Sequencing results of I-FEH I gDNA EcoTILLING fragments obtained from chicory lines susceptible (DS) or resistant (DR) to post-harvest inulin depolymerization. Partial genomic sequences are compared to a previously published 1-FEH I cDNA sequence (AJ242538). Amplification product is located inside a single exon. The cds sequences are presented in yellow. Silent SNP positions have a white background while non-silent mutations have green background. The boxes surrounding SNP positions in the cds indicate the reading frame. Positions of SNPs identified in the EcoTILLING profiles as being significantly associated are pinpointed by a red star and the estimated size of the corresponding EcoTILLING fragment. Due to their close location two SNPs could correspond to the EcoTILLING fragment “76” as indicated by the two red stars. The positions of three conserved regions located in 1-FEH I are indicated on the alignment. These regions are “WINDPNG”, “FRDP” and “WECPD”. White and grey backgrounds under these tree regions delimit the codons. Inside these conserved motifs, the three catalytic residue of GH32 family protein are colored in red. Both I-FEH I alleles are 651 bp long. From the alignment results, it appears that the allele in DS lines fully match the AJ242538 sequence while the allele in DR lines present several SNP mutations, some of them generating non-silent mutations. None of the SNP is located in any of the three conserved regions, but four SNPs induce changes in amino acid sequence and are closely located to the WECPD motif. 3D homology modeling could not identify any change in position or orientation of the E residue between the two alleles (not shown).

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest
inulin depolymerization

			1441		1500
1-FEH IIa (DR) - JQ585638 (1)			-----		GTT
1-FEH IIa (DS) - JQ585639 (1)			-----		GTT
1-FEH IIa - AJ295033 (1441)			GTGCTCATGTGTAGTGTATCTTAGCAGGTCCTACTGTTAGAAAGCAATATCGACACGACAA		GTT
Consensus (1441)					GTT
			1501		1560
1-FEH IIa (DR) - JQ585638 (4)			ATGGTGCATTGTTTGACATAGATCCTCGGCTCGAGGAGATCTCGCTAAGAACTTTGGTAT		
1-FEH IIa (DS) - JQ585639 (4)			ATGGTGCATTGTTTGACATAGATCCTCGGCTCGAGGAGATCTCGCTAAGAACTTTGGTAT		
1-FEH IIa - AJ295033 (1501)			ATGGTGCATTGTTTGACATAGATCCTCGGCTCGAGGAGATCTCGCTAAGAACTTTXXXXX		
Consensus (1501)			ATGGTGCATTGTTTGACATAGATCCTCGGCTCGAGGAGATCTCGCTAAGAACTTTGGTAT		
			1561		1620
1-FEH IIa (DR) - JQ585638 (64)			GTAATATTTTATTAAATTCATTGGTTAATGTGTATAAAGCCCTCCAAAAATTTTCATTT		
1-FEH IIa (DS) - JQ585639 (64)			GTAATATTTTATTAAATTCATTGGTTAATGTGTATAAAGCCCTCCAAAAATTTTCATTT		
1-FEH IIa - AJ295033 (1556)			XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
Consensus (1561)			GTAATATTTTATTAAATTCATTGGTTAATGTGTATAAAGCCCTCCAAAAATTTTCATTT		
			1621		1680
1-FEH IIa (DR) - JQ585638 (124)			TCACACTTTGACCCCTTGGCTATTGAATGTTTACAGATTGACCACTCAATTATTGAGAGT		
1-FEH IIa (DS) - JQ585639 (124)			TCACACTTTGACCCCTTGGCTACTGAATGTTTACAGATTGACCACTCAATTATTGAGAGT		
1-FEH IIa - AJ295033 (1556)			XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXGATTGACCACTCAATTATTGAGAGT		
Consensus (1621)			TCACACTTTGACCCCTTGGCTA TGAATGTTTACAGATTGACCACTCAATTATTGAGAGT		
			1681		1740
1-FEH IIa (DR) - JQ585638 (184)			TTCCGGAATGTTGGGAAAACATGTATCACAAGTCGGATTATACCCAAAATTTGTTAAACAAT		
1-FEH IIa (DS) - JQ585639 (184)			TTCCGGAATGTTGGGAAAACATGTATCACAAGTCGGATTATACCCAAAATTTGTTAAACAAT		
1-FEH IIa - AJ295033 (1581)			TTCCGGAATGTTGGGAAAACATGTATCACAAGTCGGATTATACCCAAAATTTGTTAAACAAT		
Consensus (1681)			TTCCGAGCTGGTGGGAAAACATGTATCACAAGTCGGATTATACCCAAAATTTGTTAAACAAT		
			1741		1800
1-FEH IIa (DR) - JQ585638 (244)			GAAGAGGCTCATCTTTTGTGTTTATAATGGGACTCAAAATGTCAAAATTTTCGAGATG		
1-FEH IIa (DS) - JQ585639 (244)			GAAGAGGCTCATCTTTTGTGTTTATAATGGGACTCAAAATGTCAAAATTTTCGAGATG		
1-FEH IIa - AJ295033 (1641)			GAAGAGGCTCATCTTTTGTGTTTATAATGGGACTCAAAATGTCAAAATTTTCGAGATG		
Consensus (1741)			GAAGAGGCTCATCTTTTGTGTTTAAATAGGGACTCAAAATGTCAAAATTTTCGAGATG		
			1801		1860
1-FEH IIa (DR) - JQ585638 (304)			AGTGCTTGGAGTATGAAGAAATGCTAAATTTGTTGTTGATCAGAGTGTA AAAAGTGCAGCA		
1-FEH IIa (DS) - JQ585639 (304)			AGTGCTTGGAGTATGAAGAAATGCTAAATTTGTTGTTGATCAGAGTGTA AAAAGTGCAGCA		
1-FEH IIa - AJ295033 (1701)			AGTGCTTGGAGTATGAAGAAATGCTAAATTTGTTGTTGATCAGAGTGTA AAAAGTGCAGCA		
Consensus (1801)			AGTGCTTGGAGTATGAAGAAATGCTAAATTTGTTGTTGATCAGAGTGTAAAAAGTGCAGCA		
			1861		1920
1-FEH IIa (DR) - JQ585638 (364)			TAAAAAAAATAGAGGCAAGAATGTTAAAAAT-----		
1-FEH IIa (DS) - JQ585639 (364)			TAAAAAAAAT-AGAGGGCAAGAATGTTAAAAATGTAAAAAGTGCAGCATAAAAAAAATAG		
1-FEH IIa - AJ295033 (1761)			TAAAAAAAAT-AGAGGGCAAGAATGTTAAAAATGTAAAAAGTGCAGCATAAAAAAAATAG		
Consensus (1861)			TAAAAAAAAT AGAGGGCAAGAATGTTAAAAATGTAAAAAGTGCAGCATAAAAAAAATAG		
			1921		1980
1-FEH IIa (DR) - JQ585638 (397)			-----AGGTAGAGTATTCTTGTGTATAAGTAAAGT-----		
1-FEH IIa (DS) - JQ585639 (423)			AGGGCAAGAATGTTAAAAATAGGTAGAGTATTCTTGTGTATAAGTAAAGT-----		
1-FEH IIa - AJ295033 (1820)			AGGGCAAGAATGTTAAAAATAGGTAGAGTATTCTTGTGTATAAGTAAAGTAAAGGAACCA		
Consensus (1921)			AGGGCAAGAATGTTAAAAATAGGTAGAGTATTCTTGTGTATAAGTAAAGT		
			1981		2040
1-FEH IIa (DR) - JQ585638 (427)			-----		
1-FEH IIa (DS) - JQ585639 (473)			-----		
1-FEH IIa - AJ295033 (1880)			GTGAAGGTTGAAAGAGTAGGTTTGGTAGCTCTTTCGTAAAGATCATCTTTGTTGCTGTGA		
Consensus (1981)					
			2041		2100
1-FEH IIa (DR) - JQ585638 (427)			-----		
1-FEH IIa (DS) - JQ585639 (473)			-----		
1-FEH IIa - AJ295033 (1940)			TTCTTGTGTAAAAATTTTGAGGTTGTATCGATTTTTATAGCCTTTGTAAGTTATTGAGA		
Consensus (2041)					
			2101		2160

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest
inulin depolymerization

```

1-FEH IIa (DR) - JQ585638(427) -----
1-FEH IIa (DS) - JQ585639(473) -----
1-FEH IIa - AJ295033 (2000) TGTAAGATTGTGCAAAGCATTGTAACAAATTATTGTTTCATTAATACAACATACATATT
      Consensus (2101)
                                2161                                2187
1-FEH IIa (DR) - JQ585638(427) -----
1-FEH IIa (DS) - JQ585639(473) -----
1-FEH IIa - AJ295033 (2060) TGAGCAAAAAAAAAAAAAAAAAAAAAA
      Consensus (2161)

```

Supplementary material 4: Sequencing results of I-FEH IIa gDNA EcoTILLING fragments obtained from chicory lines susceptible (DS) or resistant (DR) to post-harvest inulin depolymerization. Partial genomic sequences are compared to a cDNA sequence (AJ295033). The location of a 100 bp intron is indicated by “x” in AJ295033. Cds sequences are presented with a yellow background while intron and 3’UTR are identified by a grey background. Silent SNP positions have a white background and non-silent mutations are colored in green. The boxes surrounding SNP positions in the cds indicate the reading frame. Non-coding regions have a grey background and red font. In these regions, SNP loci are identified by black font. Position of SNP mutations identified in the EcoTILLING profile as being significantly associated are pinpointed with a red star and the estimated sizes of the corresponding EcoTILLING fragments are indicated. The terminal stop codon has a red background. I-FEH IIa (DS) and (DR) are respectively 472 and 426 bp long with a 1 bp indel located at the beginning of the 3’UTR and a second larger indel of 47 bp. The susceptible individuals possess two perfect tandem copies of this 47 bp region. This 47 bp sequence comprise the six last codons of the cds and the 29 first bp of the 3’UTRand is identical to the 3’UTR of 1-FEH IIa - AJ295033 while the partial 3’UTR of the “resistant” samples present two SNPs in the sequence corresponding to this 47 bp repeat. Most SNPs identified in the cds are silent mutations. The allele identified in DS lines, containing the duplication are identified in the manuscript as the D+ allele, while the 1-FEH IIa allele presenting no duplication, identified in DR lines, will be referred to in the manuscript as de d- allele.

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest
inulin depolymerization

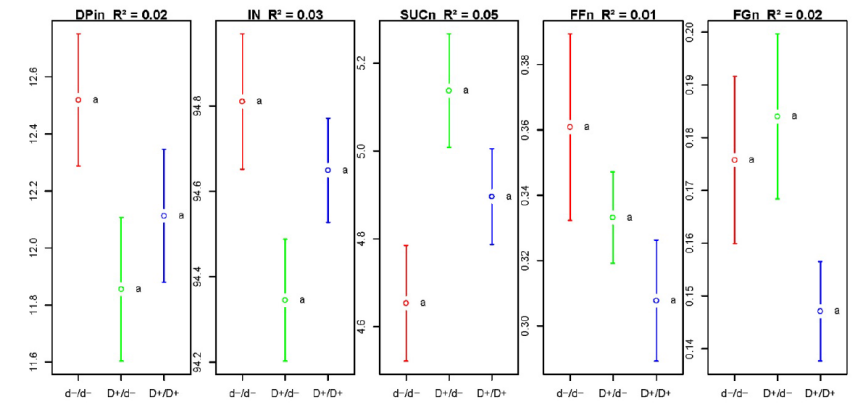
		1441		1500
1-FEH IIb (DR) - JQ585641 (1)		-----	GTATGGTGCATTGTTGAC	
1-FEH IIb (DS) - JQ585640 (1)		-----	GTATGGTGCATTGTTGAC	
1-FEH IIb - AJ295034 (1441)		CTTAGCAGGTCTACCGTTAGGAGCAACATCGACACGACAA	GTATGGTGCATTGTTGAC	
Consensus (1441)			GTATGGTGCATTGTTGAC	
		1501		1560
1-FEH IIb (DR) - JQ585641 (21)		ATAGATCCTTAAGTCGGAGGAGATCTCGCTTAGAAACTTGGTATGGAATGTTTTATTAGT		
1-FEH IIb (DS) - JQ585640 (21)		ATAGATCCTTAAGTCGGAGGAGATCTCGCTTAGAAACTTGGTATGGAATGTTTTATTAGT		
1-FEH IIb - AJ295034 (1539)		ATAGATCCTTAAGTCGGAGGAGATCTCGCTTAGAAACTTGGTATGGAATGTTTTATTAGT		
Consensus (1501)		ATAGATCCTTAAGTCGGAGGAGATCTCGCTTAGAAACTTGGTATGGAATGTTTTATTAGT		
		1561		1620
1-FEH IIb (DR) - JQ585641 (81)		GTCATGGATGCCCTCTAAATTTCCATCTTTACACTTTTGACCCCTTGTGTTGTAATGT		
1-FEH IIb (DS) - JQ585640 (81)		GTCATGGATGCCCTCTGAAA-TTTCATCTTTACACTTTTGACCCCTTGTGTTGTAATGT		
1-FEH IIb - AJ295034 (1539)		XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX		
Consensus (1561)		GTCATGGATGCCCTCT AAA TTTCCATCTTTACACTTTTGACCCCTTGTGTTGTAATGT		
		1621		1680
1-FEH IIb (DR) - JQ585641 (141)		TTACAGATAGACCACTCAATTATTGAGAGTTTCGAGCGGGAAGGACATGTATCAC		
1-FEH IIb (DS) - JQ585640 (140)		TTATAGATAGACCACTCAATTATTGAGAGTTTCGAGCGGGAAGGACATGTATCAC		
1-FEH IIb - AJ295034 (1539)		XXXXXXXXGATAGACCACTCAATTATTGAGAGTTTCGAGCGGGAAGGACATGTATCAC		
Consensus (1621)		TTA AGATAGACCACTCAATTATTGAGAGTTTCGAGCGGGAAGGACATGTATCAC		
		1681		1740
1-FEH IIb (DR) - JQ585641 (201)		AGTCGGATTACCCACAATTGTGTAACAATGAAGATCTCATCTTTTTCGCTTTAATAAC		
1-FEH IIb (DS) - JQ585640 (200)		AGTCGGATTACCCACAATTGTGTAACAATGAAGATCTCATCTTTTTCGCTTTAATAAC		
1-FEH IIb - AJ295034 (1594)		AGTCGGATTACCCACAATTGTGTAACAATGAAGATCTCATCTTTTTCGCTTTAATAAC		
Consensus (1681)		AGTCGGATTACCCACAATTGTGTAACAATGAAGATCTCATCTTTTTCGCTTTAATAAC		
		1741		1800
1-FEH IIb (DR) - JQ585641 (261)		GGGACTCAAAATGTTAAAATTTCGAAGATGAGTGCTTGGAGTATGAAGAATGCTGCATT		
1-FEH IIb (DS) - JQ585640 (260)		GGGACTCAAAATGTTAAAATTTCGAAGATGAGTGCTTGGAGTATGAAGAATGCTGCATT		
1-FEH IIb - AJ295034 (1654)		GGGACTCAAAATGTTAAAATTTCGAAGATGAGTGCTTGGAGTATGAAGAATGCTGCATT		
Consensus (1741)		GGGACTCAAAATGTTAAAATTTCGAAGATGAGTGCTTGGAGTATGAAGAATGCTGCATT		
		1801		1860
1-FEH IIb (DR) - JQ585641 (321)		GTGTTGATCAGCAATTTAAAAAGTGCAGCAATAAAATATTAGAGGGGCTAGAATGTTAAA		
1-FEH IIb (DS) - JQ585640 (320)		GTGTTGATCAGCAATTTAAAAAGTGCAGCAATAAAATATTAGAGGGGCTAGAATGTTAAA		
1-FEH IIb - AJ295034 (1714)		GTGTTGATCAGCAATTTAAAAAGTGCAGCAATAAAATATTAGAGGGGCTAGAATGTTAAA		
Consensus (1801)		GTGTTGATCAGCAATTTAAAAAGTGCAGCAATAAAATATTAGAGGGGCTAGAATGTTAAA		
		1861		1920
1-FEH IIb (DR) - JQ585641 (381)		AATAGATATTCTTGTATAAAGTAAAGTAAAGGAAGCAGTGAAGGTCGAAAGAGTAGGTT		
1-FEH IIb (DS) - JQ585640 (380)		AATAGATATTCTTGTATAAAGTAAAGTAAAGGAAGCAGTGAAGGTCGAAAGAGTAGGTT		
1-FEH IIb - AJ295034 (1774)		AATAGATATTCTTGTATAAAGTAAAGTAAAGGAAGCAGTGAAGGTCGAAAGAGTAGGTT		
Consensus (1861)		AATAGATATTCTTGTATAAAGTAAAGTAAAGGAAGCAGTGAAGGTCGAAAGAGTAGGTT		
		1921		1980
1-FEH IIb (DR) - JQ585641 (441)		TGGTAGCTCTTTCATAAAGATCATTCTTTGTGCTTTATTCTTGTGTAAAAATTTTTCG		
1-FEH IIb (DS) - JQ585640 (440)		TGGTAGCTCTTTCATAAAGATCATTCTTTGTGCTTTATTCTTGTGTAAAAATTTTTCG		
1-FEH IIb - AJ295034 (1834)		TGGTAGCTCTTTCATAAAGATCATTCTTTGTGCTTTATTCTTGTGTAAAAATTTTTCG		
Consensus (1921)		TGGTAGCTCTTTCATAAAGATCATTCTTTGTGCTTTATTCTTGTGTAAAAATTTTTCG		
		1981	94 *	2040
1-FEH IIb (DR) - JQ585641 (501)		GTGTTATCGACCTTTTGTGCATTATAAGTTATCGAGATGAAGATTTTGTGCAAGAC		
1-FEH IIb (DS) - JQ585640 (500)		GTGTTATCGACCTTTTGTGCATTATAAGTTATCGAGATGAAGATTTTGTGCAAGAC		
1-FEH IIb - AJ295034 (1894)		GTGTTATCGACCTTTTGTGCATTATAAGTTATCGAGATGAAGATTTTGTGCAAGAC		
Consensus (1981)		GTGTTATCGACCTTTTGTGCATTATAAGTTATCGAG		
		2041		2100
1-FEH IIb (DR) - JQ585641 (561)		ATGAATGCCATGATGCCATTGTAAC-----		
1-FEH IIb (DS) - JQ585640 (560)		ATGAATGCCATGATGCCATTGTAAC-----		
1-FEH IIb - AJ295034 (1954)		ATGAATGCCATGATGCCATTGTAACAAATTATTGTTCCATTAAAAAAAAAAAAAAAAAA		
Consensus (•2018)				

```

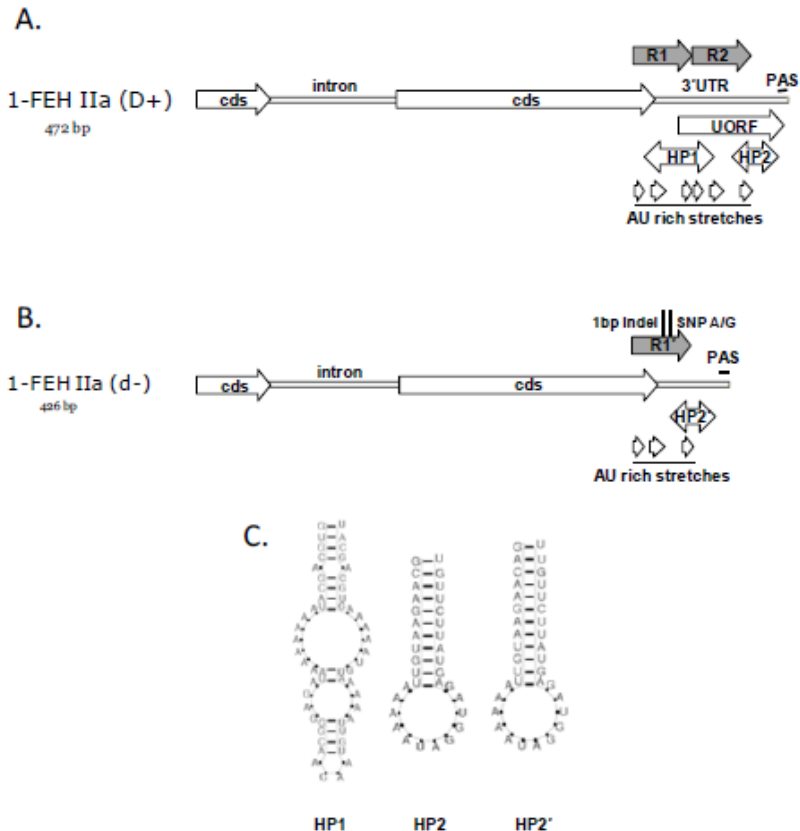
                                2101
1-FEH IIb (DR) - JQ585641(586) ---
1-FEH IIb (DS) - JQ585640(585) ---
1-FEH IIb - AJ295034 (2014) AAA
          Consensus (•2018)

```

Supplementary material 5: Sequencing results of I-FEH IIb gDNA EcoTILLING fragments obtained from chicory lines susceptible (DS) or resistant (DR) to post-harvest inulin depolymerization. Partial genomic sequences are compared to a published cDNA sequence (AJ295034). The location of a 86/87 bp intron is indicated by “x” in AJ295034. Cds sequences are presented with a yellow background while intron and 3’UTR are identified by a grey background. Silent SNP positions have a white background and non-silent mutations have a green one. The boxes surrounding SNP positions in the cds indicate the reading frame. Non-coding regions have a grey background and red font. In these regions, SNP loci are identified by black font. Position of SNP mutations identified in the EcoTILLING profile as being significantly associated are pinpointed by a red star and the estimated size of the corresponding EcoTILLING fragment is indicated. The terminal stop codon has a red background. I-FEH IIb (DS) and (DR) are respectively 584 and 585 bp long. Only two of the 10 SNP positions induce non-silent mutations. A one bp indel was also detected in the last intron.



Supplementary material 6: Means and standard error of the three different 1-FEH IIa genotypes (“d-“ refers to the allele ”absence of duplication” and the “D+” to the allele “presence of the duplication”) among 116 lines for five carbohydrates related parameters before exposure to cold temperatures: inulin degree of polymerization (DPIn), inulin content (IN), sucrose content (SUCn), free fructose content (FFn) and free glucose content (FGn). No statistical differences are observed.



Supplementary material 7: Schematic representation of the two 1-FEH IIa alleles respectively identified in depolymerization susceptible (DS lines, D+ allele, JQ585639) or depolymerization resistant (DR lines, d- allele, JQ585638) lines. The two alleles mainly differ by the presence of a perfect tandem repeat (D+) or by a single copy of this repeated region (d-) (A-B). An indel is located in the 3'end of the cDNA. This 47 bp long repeat starts near the 3'end of the cds and terminates in the very first part of the 3'UTR. D+ exhibit a longer 3'UTR containing a perfect tandem repeat of the 47 bp motif (R1 and R2) while the corresponding R1' motif in d- presents a one bp indel as well as one SNP polymorphism when compared to R1. (C) RNApromo website identified a conserved hairpin structure present in the two alleles (HP2 and HP2') as well as another secondary structure induced by the tandem repetition (HP1). A scan of the last exon against the UTRdb identified a

Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest
inulin depolymerization

UORF region in the 3'UTR of D+, while the sequence is absent from d-. Last structures identified are AU rich stretches. These stretches were not detected as ARE, but are enriched in the 3'UTR of D+.

7 Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in Cichorium intybus than copy number variation of its close paralog 1-FEH IIa.

This is an adapted version of the research article entitled:

Dauchot, N., Raulier, P., Maudoux, O., Notté, C., Draye, X., & Van Cutsem, P. (2015). Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in Cichorium intybus than copy number variation of its close paralog 1-FEH IIa. *Frontiers in Plant Science*, 6, 455.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

7.3 Abstract

Inulin polyfructan is the second most abundant storage carbohydrate in flowering plants. After harvest, it is depolymerized by fructan exohydrolases (FEHs) as an adaptive response to end-season cold temperatures. In chicory, the intensity of this depolymerization differs between cultivars but also between individuals within a cultivar. Regarding this phenotypic variability, we recently identified statistically significant associations between inulin degradation and genetic polymorphisms located in three fructan exohydrolases. We present here new results of a systematic analysis of copy number variation (CNV) in five key members of the chicory (*Cichorium intybus*) GH32 multigenic family, including three FEH genes and the two inulin biosynthesis genes: 1-SST and 1-FFT. qPCR analysis identified a significant variability of relative copy number only in the 1-FEH IIa gene. However, this CNV had no quantitative effect. Instead, cloning of the full length gDNA of a close paralogous sequence (1-FEH IIb) identified a 1028 bp deletion in lines less susceptible to post-harvest inulin depolymerization. This region comprises a 9 bp mini-exon containing one of the three conserved residues of the active site. This results in a putative non-functional 1-FEH IIb allele and an observed lower inulin depolymerization. Extensive genotyping confirmed that the loss of mini-exon 2 in 1-FEH IIb and the previously identified 47 bp duplication located in the 3'UTR of 1-FEH IIa belong to a single haplotype, both being statistically associated with reduced susceptibility to post-harvest inulin depolymerization. Emergence of these haplotypes is discussed.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

7.4 Introduction

Industrial chicory is the main commercial source of inulin, a linear fructose polymer used by the agro-industry as texturizer, fat substitute (Mendoza et al., 2001; Keenan et al., 2014; Karimi et al., 2015) or low calories sweetener. Inulin is also used by pharmaceutical industries for its health promoting properties (Gibson et al., 2006; Meyer and Stasse-Wolthuis, 2009). In chicory, like in most Asteraceae (Hendry, 1993), inulin is the main reserve carbohydrate. It is accumulated in the taproot during growing season and it is hydrolyzed by fructan exohydrolases (FEHs) as an adaptive response to the exposure to end-season cold temperatures (Dauchot et al., 2014).

As first proposed in 1968 in *Helianthus tuberosus* (Edelman and Jefford, 1968), inulin synthesis results from the sequential action of the enzyme sucrose:sucrose 1-fructosyltransferase (1-SST) and fructan:fructan 1-fructosyltransferase (1-FFT). The 1-SST enzyme initiates inulin synthesis by transferring a fructose moiety from a sucrose donor molecule (GF) to a second sucrose molecule, which acts as fructose acceptor, to produce 1-kestotriose (GFF) and free glucose (G). This 1-kestotriose molecule is then used as fructose donor by the enzyme fructan:fructan 1-fructosyltransferase (1-FFT) to elongate inulin molecules (GFF_n+1), releasing free sucrose (GF). Several years later, we were able to isolate the corresponding 1-FFT and 1-SST coding sequences from chicory (de Halleux and Van Cutsem, 1997). The coding sequences of chicory 1-FEH I, 1-FEH IIa and 1-FEH IIb inulin hydrolyzing genes were isolated short after (Van den Ende et al., 2000; Van den Ende et al., 2001).

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

In chicory, inulin chain length varies between 10 and up to 60 fructosyl units. Average inulin chain length varies along the season. At the end of the growing season, inulin is hydrolyzed to help face winter cold temperatures, resulting in an increase of free sugars (sucrose, glucose and fructose) and a global shortening of inulin chain length. Since chicory 1-SST and FEH genes are reported to be mainly regulated at the transcriptional level (Michiels et al., 2004), the reduction of the average inulin chain length and parallel increase of free sugars is believed to result from a reduction of 1-SST transcription and a parallel increase of the transcription of genes coding for inulin hydrolyzing enzymes (van Arkel et al., 2012), namely FEHs, while the expression of 1-FFT remains constant (Van Laere and Van den Ende, 2002).

This adaptive response to cold seriously interferes with the industrial inulin extraction process because roots start to accumulate free sugars, which negatively affects the inulin extraction process and final yield. The average size reduction of the inulin molecules also modifies their physico-chemical properties (Mensink et al., 2015) and commercial value.

Inulin chain length differs between species (Itaya et al., 2007). While the inter-specific variability of inulin chain length is believed to result from a difference of activity of the enzyme 1-FFT (Hellwege et al., 1998), the molecular bases of the high intra-specific variation of inulin chain length observed in industrial chicory cultivars (Amaducci and Pritoni, 1998; Wilson et al., 2004; Monti et al., 2005) are still unknown.

In this context, we sampled the existing phenotypic diversity observed between

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

and within industrial chicory cultivars to identify the genetic bases responsible for the observed differences of susceptibility to post-harvest inulin depolymerization. In a previous study (Dauchot et al., 2014), we identified SNP polymorphisms statistically associated with these differences of susceptibility. These SNPs were located within a single subgroup of the GH32 multigenic family, namely the FEHs 1-FEH I, 1-FEH IIa and 1-FEH IIb, a class of enzymes that specifically degrades inulin (Van den Ende et al., 2000; Van den Ende et al., 2001). To further investigate the genetic bases of the observed phenotypic plasticity, we decided to study copy number variation (CNV) of these three FEHs and of two closely related GH32 enzymes involved in inulin biosynthesis (1-SST and 1-FFT).

CNV are extensively studied in humans and have been associated with several diseases (Almal and Padh, 2011). CNV have only recently been recognized as responsible for the high phenotypic plasticity observed in domestic dog (Alvarez and Akey, 2011; Nicholas et al., 2011) and other domestic animal species (Clop et al., 2012). In crop species, CNV started not long ago to be the subject of a growing interest in the scientific community (Springer et al., 2009; Knox et al., 2010). CNV is a spontaneous process whose rate could exceed the rate of single nucleotide polymorphisms by two orders of magnitude, as observed in *Caenorhabditis* (Lipinski et al., 2011). In *Arabidopsis*, CNV has been observed over immediate family generational scales (DeBolt, 2010). This author highlighted a putative effect of the environment on CNV and also pointed out that tandem-duplicated genes were common in CNV events. The sequence homology (94%) between the 1-FEH IIa (AJ295033) and 1-FEH IIb (AJ295034) genes coupled to their physical proximity (Cadalen et al., 2010) could favor

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

CNV appearance in this multigenic family and result in the observed phenotypic variability of carbohydrate content and properties in chicory after harvest.

Here we identified the presence of CNV in the 1-FEH IIa gene. Despite putative functional relevance of these results, statistical tests confirmed that these CNV were not associated to quantitative variations of susceptibility to post-harvest inulin-depolymerization. On the other hand, we were also able to identify a loss-of-function allele in a close paralog (1-FEH IIb) resulting from the loss of one of the three amino acids present at the active site of the enzyme. This allele was associated with a reduced susceptibility to post-harvest inulin depolymerization.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

7.5 Materials and Methods

7.5.1 Plant Material

Analyses were performed on a subset of 112 industrial chicory lines originating from a collection described earlier (Dauchot et al., 2014). The 112 lines were created by selfing of 112 individual plants selected from a collection of 600 individuals randomly sampled out of 18 varieties at the origin of the selection of modern industrial chicory cultivars. The 112 individuals were selected using the Mstrat sampling strategy, based on 15 SSRs markers (Gouesnard et al., 2001) to gather maximum diversity of the original collection and minimize the structure of the sampled population.

The 112 lines were sown on 2 May 2011 in Warcoing (7740, Belgium) and 6 to 24 roots per line were harvested between Oct 2 and Oct 9, 2011. These 112 lines were used for carbohydrate analysis, while the gDNA of the corresponding parents were used for FEH genotyping and qPCR analysis.

7.5.2 Root sampling and carbohydrates phenotyping

Due to the amount of material needed for carbohydrate analysis and the destructive nature of the procedure, analyses were performed on two pools of

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

root tissues obtained after rasping of several roots from the same line. This provided an estimate of the phenotype of the 112 parental plants at two distinct time points, at harvest and after harvest respectively. The roots were phenotyped for five carbohydrate characteristics: the inulin degree of polymerization (DP_{in}), inulin content (IN), sucrose content (SUC_n), free fructose content (FF_n) and free glucose content (FG_n). The IN, SUC_n, FF_n and FG_n were expressed as a percentage of the total carbohydrate mass. Analyses were performed according to Van Waes (Van Waes et al., 1998) as described earlier (Dauchot et al., 2014).

7.5.3 Leaf sampling and DNA extraction

Genomic DNA was extracted from the parent plants of the 112 lines. Upon sampling, fresh chicory leaves were stored in paper envelopes at -80°C until further processing. For gDNA extraction, frozen leaves were ground using a MM400 mixer mill (Retsch, Düsseldorf, Germany) for 45 s at 30 Hz in 35 ml stainless steel screw-top grinding jars (Retsch, # 01.462.0214) pre-cooled in liquid nitrogen. Leaf powder was immediately returned to -80°C until extraction. 100 mg of frozen leaf powder was used as starting material for genomic DNA extraction using NucleoSpin PlantII extraction kit according to manufacturer's instructions (Macherey-Nagel, Duren, Germany). First step of the protocol uses a buffer containing RNase. Following extraction, gDNA was evaluated for integrity by agarose gel electrophoresis and for purity by spectrophotometric analysis at 260 and 280 nm on a Thermoscientific Multiskan GO device. 260/280 ratio quality threshold was set to 1.8 and could not exceed 2.1. Further validation of the quality and/or presence of PCR inhibitors was evaluated by performing a

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

classical amplification followed, in case of positive amplification, by the evaluation of qPCR efficiency of DNA pools with qPCR primer pairs.

7.5.4 PCR and qPCR primer design

qPCR primers were picked manually applying selection criteria as described earlier (D'Haene et al., 2010). Melting temperature was calculated using the ABi online tool available at <http://www6.appliedbiosystems.com/support/techtools/calc/> with a salt concentration of 43 mM and a primer concentration of 0.1 μ M. Primers used for qPCR analyses are listed in Table 13. All primers were synthesized by Integrated DNA technologies (Leuven, Belgium) and were purified by standard desalting.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Target	PCR/qPCR	GenBank	Author	Primer sequence	Tm (°C)	Efficiency	Calibration Curve	R ²	Amplicons
Actin	qPCR	EF528575	(Maroufi et al., 2010)	F: CCAAATCCAGCTCATCAGTCG R: TCTTTCGGCTCCGATGGTGAT	60	P1: 107.9 ± 3.4 P2: not tested P3: not tested	-3.1 log Q0 + 24.4	0.99	gDNA: 74 bp cDNA: 74 bp
β-tubulin	qPCR	AF101419	(Maroufi et al., 2010)	F: GCACGGCATTGATGTGACC R: GAACAAACCTCCC GCCACT	60	P1: 102.5 ± 4.3 P2: 105.3±2.1 P3: 103.5±3.2	-3.2 log Q0 + 25.6 -3.2 log Q0 + 22.7 -3.2 log Q0 + 25.4	0.99 1.00 0.99	gDNA: 101 bp cDNA: 101 bp
1-SST	qPCR	U81520	this paper	F: ATCTCCCATTCGCCATGGT R: TGCCAGTGTAGAGCATGATGATCT	60	P1: 104.2 ± 4.4 P2: 96.2±3.4 P3: 99.9±2.6	-3.2 log Q0 + 25.6 -3.4 log Q0 + 25.6 -3.3 log Q0 + 25.5	0.99 0.99 0.99	gDNA: 108 bp cDNA: 108 bp
1-FFT	qPCR	U84398	this paper	F: GAGCTTCCCGTAGCCTTGAC R: CATTTCCGGTGTAATGCA	60	P1: 103.8 ± 4.7 P2: 98.8±2.8 P3: 100.5±3.2	-3.2 log Q0 + 25.6 -3.3 log Q0 + 25.4 -3.3 log Q0 + 25.5	0.99 0.99 0.99	gDNA: 112 bp cDNA: 112 bp
1-FEH I	qPCR	AJ242538	this paper	F: TTTCTTCTCGAACCAGCTCTCAG R: CGTCTTGACCGGTATATAGAATTATGG	60	P1: 102.0 ± 2.1 P2: 99.1±2.6 P3: 99.9±2.6	-3.2 log Q0 + 22.7 -3.3 log Q0 + 22.3 -3.3 log Q0 + 22.6	0.99 0.99 0.99	gDNA: 118 bp cDNA: 118 bp
1-FEH IIa	qPCR	JQ585639	this paper	F: TCACACTTTGACCCCTTGGCTA R: GGGTAAATCCGACTTGTGATACATGTT	60	P1: 102.1 ± 3.3 P2: 99.4±2.3 P3: 98.0±3.7	-3.2 log Q0 + 23.6 -3.3 log Q0 + 23.2 -3.3 log Q0 + 23.4	0.99 0.99 0.99	gDNA: 103 bp cDNA: none
1-FEH IIb	qPCR	JQ585640	this paper	F: TCTTTACACTTTTGACCCCTTGTGTTG R: GGGTAAATCCGACTTGTGATACATGTC	60	P1: 99.7 ± 4.5 P2: 97.5±3.0 P3: 98.9±4.	-3.3 log Q0 + 24.7 -3.3 log Q0 + 24.3 -3.3 log Q0 + 24.5	0.99 0.99 0.99	gDNA: 105 bp cDNA: none

Table 13: qPCR amplification primers, Tm and calculated efficiencies. According to plate design constraints, the 36 samples of this study were analyzed in three sets (P1, P2 and P3). For each set, gDNAs were pooled and used to estimate PCR efficiencies of five target genes and two normalization genes (Actin and β-tubulin). P1 was used for first tests. Based on first test (P1), β-tubulin was preferred as normalization gene.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Prior to qPCR analysis, all gDNA were tested for amplifiable RNA/cDNA contaminants with a primer pair targeting a genomic region including an intron. All samples were checked for single band amplification profiles. The primer pair used for this test targets 1-FEH I (AJ2425378) and should amplify a 891 bp genomic fragment and only 395 bp on cDNA. Primer sequences were F- GCACTTTTCTAGTAAAACGGG and R- TCCGGTTATTGCTAAGCCAG. Amplifications were typically performed for 35 cycles with a T_m set to 55°C.

7.5.5 qPCR reaction mixture and cycling conditions

qPCR reactions were performed in a 20 μ l volume and were typically composed of 10 μ l 2x GoTaq qPCR Master Mix (Promega, #A6001), 0.2 μ l CRX internal reference (Promega, #A6001), 0.05 μ l of 100 μ M of each primer, gDNA and nuclease-free water to reach 20 μ l. Reactions were all performed in ABi StepOnePlus instrument (96 wells) with $\Delta\Delta CT$ quantification method, SYBR green amplicon detection and Standard amplification conditions consisting of an initial denaturation step at 95°C for 10 min followed by 40 cycles at 95°C for 15 sec and 60°C for 1 min. Amplification step was followed by a melting curve analysis between 60 and 95°C.

7.5.6 qPCR efficiency

According to plate design constraints, gDNA were analysed as three distinct

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

DNA sets including intra- and inter-plate normalization reference samples. PCR efficiencies were calculated for each sample set treating them as three distinct pools of 16 to 21 individual gDNA whose concentration was normalized according to spectrophotometric quantification results. For each primer pair, efficiencies were determined based on 6 gDNA concentrations of 32, 8, 2, 0.5, 0.25 and 0.125 ng/μl. Serial dilution were performed using 5 ng/μl baker yeast t-RNA carrier solution (Ambion AM7119). For each target amplicon, two “no template control” (NTC) were also included: NTC1 used nuclease-free water as template, while NTC2 used t-RNA carrier solution as template. All tests were performed on the same qPCR plate in duplicate. Each 20 μl reaction contained 10 μl of 2x GoTaq qPCR Master Mix (Promega, #A6001), 0.2 μl of CRX internal dye (Promega, #A6001), 0.05 μl of each primer at initial concentration of 100 μM, 4.4 μl of nuclease-free water and 5.3 μl of the diluted gDNA. qPCR were performed on ABi StepOnePlus™ Real-Time PCR Systems with standard 2 h runs, comparative Ct quantification ($\Delta\Delta C_t$) and SYBR® green detection parameters. PCR efficiencies were calculated with pyQPCR (available for download at <http://pyqpcr.sourceforge.net/?static2/download>), an interface for the qBase algorithm (Hellemans et al., 2007). PCR efficiency was calculated with the formula $E = (10^{(-1/\text{slope})} - 1) * 100$. The slope of the linear regression was obtained when plotting CT value vs log of the theoretical copy number. Primer pairs were considered valid when calculated efficiency was between 90 and 110% with 100% as an optimum. Standard error was calculated with pyQPCR.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

7.5.7 qPCR

Normalized relative copy number was estimated on 36 samples. Due to plate design constraints, these samples were analyzed and treated as three distinct sample sets. To allow integration of the results, an identical normalization sample (X129) was used in the three pools. Three samples were also present across plates to evaluate the validity of the normalization and the inter-run reproducibility.

The three sample sets were extracted with the same protocol and followed the same validation pipeline. All samples were normalized to the same concentration (8 ng/ μ l) based on the results obtained during efficiency estimates in order to reach an average CT of 20. Efficiencies were evaluated individually for each target and each sample set to account for possible variability of DNA purity. Once prepared, a single gDNA dilution was performed for each sample and was used for all tests. Samples were not re-extracted or re-diluted between analyses to improve reproducibility.

Each plate included up to 22 test samples, one normalization sample common to all plates and two no template controls (NTC1 and NTC2). On a single plate, all tests were performed in duplicate. A 96 wells plate allowed the analysis of up to two target genes.

Each 20 μ l qPCR reaction contained 10 μ l of 2x GoTaq qPCR Master Mix (Promega, #A6001), 0.2 μ l of CRX internal dye (Promega, #A6001), 0.05 μ l of each primer at initial concentration of 100 μ M, 4.7 μ l of nuclease-free water

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

and 5 µl of 8 ng/µl gDNA (40 ng).

Amplification cycles were terminated by a melting curve analysis.

7.5.8 Cloning and sequencing of 1-FEH I and 1-FEH IIb genomic regions

For cloning, gDNA amplifications were all performed on two lines highly contrasted for their susceptibility to post-harvest inulin depolymerization, namely X191 (considered as depolymerization resistant) and X200 (considered as depolymerization-sensitive) lines.

Partial genomic region of 1-FEH I was amplified with GoTaq Long Range PCR master mix (Promega) according to manufacturer's instruction with a T_m of 54.3°C, 35 cycles, elongation time of 13 min in 10 µl volume with 100 ng of gDNA in an Eppendorf Master Cycler Pro S. Due to the high sequence homology between members of the GH32 family, the genomic region was amplified with a single primer pair using long range polymerases to avoid posterior assembly of chimeric sequences that might result from the amplification of multiple partial sequences.

The two primers 1-FEHI-96F CATTGGGTTCTCTCTTTGC and 1-FEHI-1800-R CTAGGTTATAGTTACTAGAACATTATATG target the 96-1800 bp region of the 1-FEH I cDNA (AJ242538) which covers 98% (1672/1707 bp) of the corresponding coding CDS (61-1767). These primers amplified a single

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

gDNA fragment of 10,575 bp (KM494975).

Amplifications were performed with GoTaq Long Range PCR master mix (Promega) according to manufacturer's instruction with a T_m of 50°C, 35 cycles, elongation time of 7 min in 10 µl volumes with 100 ng of gDNA with primers 1FEH2a-28F CTTTTTCTCCATATGTTGTCG and 1FEH2a-5744R CAAGGAATACAGCAACAAAGAATG.

For both 1-FEH I and 1-FEH IIb, inserts were gel-purified with Nucleospin Gel extraction kit (Macherey Nagel), inserts were cloned in PCR4.1 Topo vector (Invitrogen) and electroporated in TOP10 *E. coli* (Invitrogen). Sanger sequencing of the inserts was performed at Macrogen (The Netherlands).

The parents of each 112 lines used in this study were first genotyped with the primer pair specifically targeting a 47 bp duplication in the 3'UTR region of 1-FEH IIa as described previously (Dauchot et al., 2014).

They were then genotyped for the loss of the mini-exon 2. Loss of mini-exon 2 was monitored with primers F2ab-638F-AGAAACTACAACTCATAAATGAATATGC and F2ab-865R-CTTTGAAATRTCTAGCCGCCGTAAAC using GoTaq Long PCR master Mix (Promega # M4021) according to manufacturer's instruction for 35x cycles at 54.3°C. This primer pair amplified a region located in the promoter of 1-FEH IIb and presented size polymorphism between the two 1-FEH IIb alleles (observed 247/302 bp) (X200 - expected 250 bp, X191 - expected 305). Due to the high homology between 1-FEH IIa and 1-FEH IIb promoters, these primers

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

also amplified 1-FEH IIa but with a distinct amplification size (observed 225/229 bp). The amplification size was in agreement with 1-FEH IIa genomic reference sequence (AY323935 - expected 228 bp).

7.5.9 Data analysis

Generation of normalized relative quantification (NRQ) values and standard error was performed using pyQPCR software as an interface to treat Abi StepOnePlus raw data with the qBase algorithm (Hellemans et al., 2007). $\Delta\Delta CT$ relative quantification was performed using β -tubulin as normalization gene.

7.5.10 Statistical tests

The association between the carbohydrate characteristics and the FEH genotype with the different NRQ values was analyzed respectively by ANOVA or linear regression, using the aov function of R version 3.0.0 (R Development Core Team, 2013). Model comparison was realized with the anova functions of R version 3.0.0 (R Development Core Team, 2013).

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

7.6 Results

7.6.1 Normalization genes for qPCR in chicory

This paper is, to our knowledge, the first one describing CNV analysis by qPCR in chicory. As a result, no normalization genes have been reported for this purpose so far. A paper reporting the validation of reference genes for real-time gene expression analysis (RT-PCR) in chicory evaluated the expression stability of nicotinamide adenine dinucleotide dehydrogenase (NADHD), actin (ACT), α -tubulin (TUB), glyceraldehyde 3-phosphate-dehydrogenase (GAPDH), histone H3 (H3), elongation factor 1-alpha (EF) and 18S rRNA (rRNA) and concluded that ACT, EF, rRNA, H3 and TUB could be used for normalization (Maroufi et al., 2010). Out of these genes, we tested ACT and TUB as normalization genes for CNV analysis.

Our results indicate that, based on a preliminary analysis performed on 20 individuals (Pool 1, P1), from our chicory sampling population, estimates of relative copy number of 1-FFT gene using actin as normalization gene resulted in a higher variability and standard error deviation than when normalized against β -tubulin. In the initial test, efficiency of β -tubulin was also closer to the 100% optimum (102.57 ± 4.37) as compared to actin (107.94 ± 3.48). β -tubulin was therefore used as a quantification reference. The apparent stability of 1-FFT copy number makes it a good candidate as an additional reference gene for copy number analysis in chicory.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

7.6.2 Specificity of qPCR primer pairs

For all tests, terminal melting curve analysis never identified any multiple melting temperatures. Specificity of actin and tubulin primers was validated previously (Maroufi et al., 2010). The actin qPCR primer pair was valid for use on cDNA as well as on gDNA (same amplification size). Tubulin qPCR primer pair was also located inside a single exon according to gDNA amplification size and personal unpublished gDNA sequences. 1-SST, 1-FFT and 1-FEH I specific primer pairs were all located within a single exon. 1-FEH IIa and 1-FEH IIb are two paralog sequences presenting more than 94% homology of their coding sequence. Based on four partial gDNA sequences published recently (JQ585641, JQ585640, JQ585639 and JQ585638), highly specific primer pairs could be designed to amplify a region covering part of the last intron and part of the last exon. This design ensured a high specificity for qPCR use. All primer pairs and accession numbers of their respective target sequences are listed in Table 13.

7.6.3 qPCR efficiencies

Efficiencies of qPCR primer pairs were calculated as explained in the material and methods section. Results are summarized in Table 13. Efficiencies of all primer pairs included in this study were between 96.2% and 105.4%. Efficiencies were estimated for each gDNA set independently to account for

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

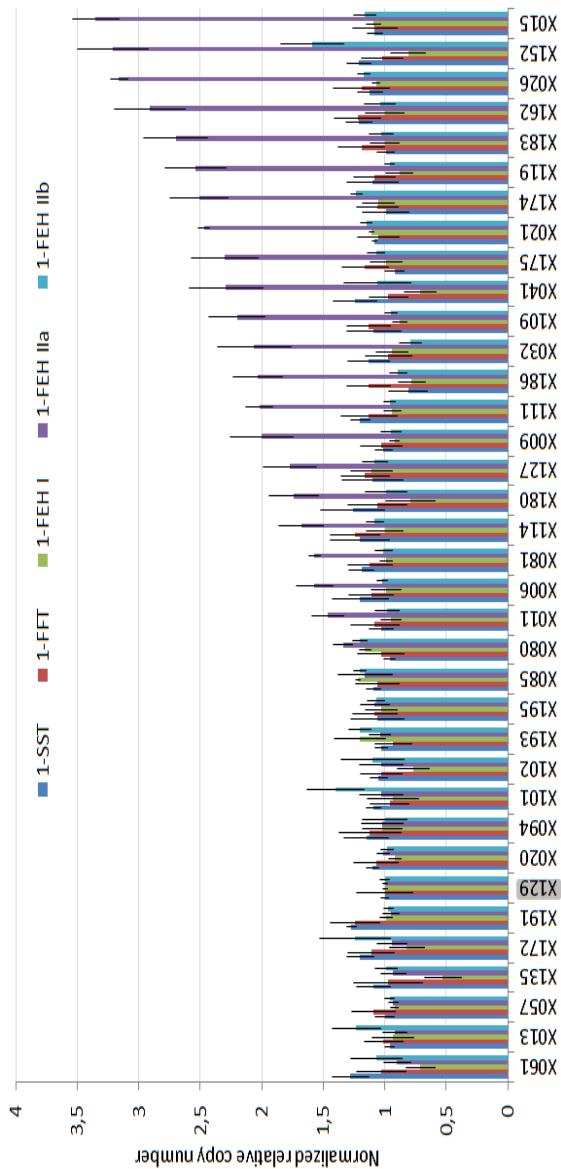
variability between the investigated gDNA sets.

7.6.4 Association of GH32 NRQ data with post-harvest inulin depolymerization

We investigated the relative copy number of five members of the GH32 family by qPCR on the genomic DNA of the parents of 36 lines used for the phenotyping.

The normalized relative quantities (NRQ) values for 1-SST, 1-FFT, 1-FEH I and 1-FEHIIb showed a small variability, ranging from 0.81, 0.93, 0.53, 0.79 to 1.28, 1.24, 1.22 and 1.59 respectively, while 1-FEH IIa showed a wider range of NRQ values ranging from 0.90 up to 3.35 (Figure 40).

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.



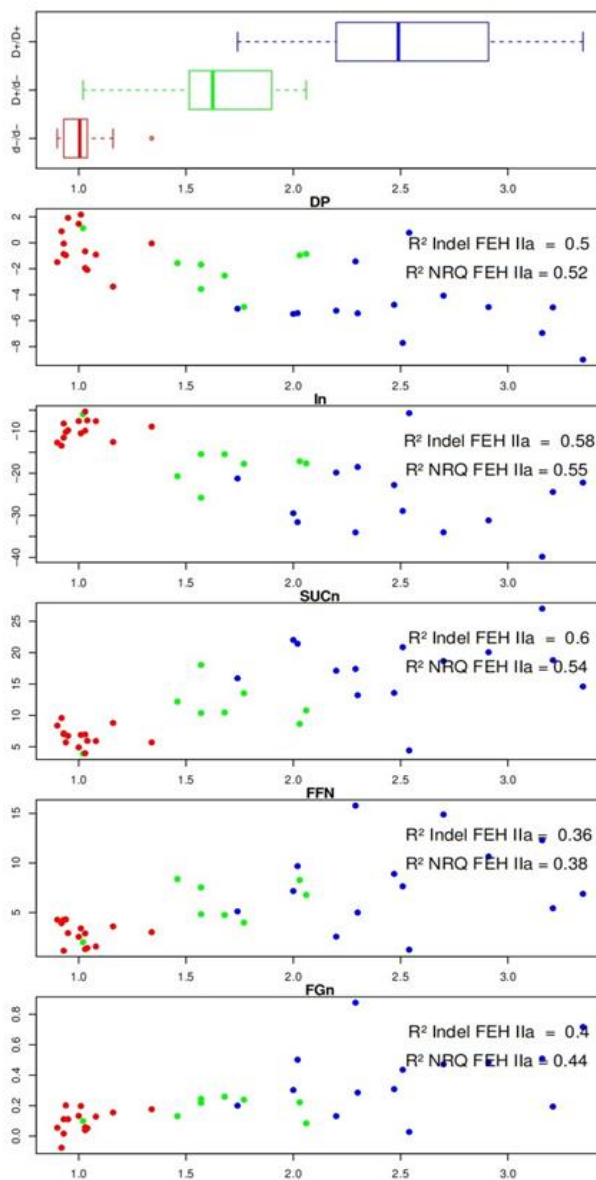
Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Figure 40: Plot of the normalized relative quantities (NRQ) and standard error estimated for 5 members of the GH32 multigenic family by qPCR on the genomic DNA of 36 chicory lines exhibiting different susceptibility to post-harvest inulin depolymerization. X129 was used to normalize the results. Significant differences of relative copy number were only detected for 1-FEH IIa.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

We performed statistical tests to evaluate the correlation between NRQ values and phenotypic data recorded at harvest and after harvest. These results are presented in Figure 41.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.



Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Figure 41: For 36 lines, boxplot of the genotype of the duplication located in 3'UTR region of 1-FEH IIa, plot of the difference of the inulin degree of polymerization (DPin), inulin content (IN), sucrose content (SUCn), free fructose content (FFn) and free glucose content (FGn) before and after exposure to post-harvest cold temperatures (Y axis) according to the FEH IIa NRQ value (X axis). Each data point is colored according to the nature of the FEH genotype (3' duplication) previously associated with contrasted susceptibilities to post-harvest inulin depolymerization. Blue: homozygous for the FEH IIa allele presenting no duplication in the 3' UTR, red: homozygous for the FEH IIa allele presenting the 47 bp duplication in its 3' UTR, green: heterozygote.

For each plot, we also included a three color code, referring to the genotype of 1-FEH IIa. We previously demonstrated a strong correlation between the presence/absence of a 47 bp duplication in the 3'UTR of 1-FEH IIa and the susceptibility to post-harvest inulin depolymerization. A graphical representation to this 47 bp duplication can be found in the fifth supplementary material of our previous chapter (Dauchot et al., 2014). Figure 41 also presents the R^2 values obtained for the presence/absence of the 47 bp long duplication in the 3'UTR of 1-FEH IIa (model 1) and for the loss of mini-exon 2 in 1-FEH IIb (model 2) respectively.

Among the 36 parents analyzed by qPCR, 15 plants were genotyped as being homozygous for the absence of duplication in the 3'UTR (red: d-/d- considered as more resistant to depolymerization), 14 plants were homozygous for the presence of duplication (blue: D+/D+ , considered as depolymerization prone) and 8 heterozygous plants (green: D+/d- presenting an intermediate phenotype).

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

For these three categories, the 1-FEH IIa NRQ values gave averages of 1.02, 1.64 and 2.5 respectively (Figure 41, upper box). The variability in 1-FEH2a relative copy number was very low for the d-/d-, higher for the D+/d- and highest for D+/D+. The genotype for the 1-FEH IIa 3' duplication was significantly correlated and explained an important proportion of variation of the 1-FEH IIa normalized relative copy number ($p\text{Val} = 3.2 \times 10^{-12}$ $R^2 = 0.78$). From these results, we conclude that the presence of a perfect 47 bp tandem duplication in the 3'UTR of 1-FEH IIa is correlated with the presence of multiple copies of 1-FEH IIa. Low 1-FEH IIa copy number is, on the other hand, correlated to the absence of the 47 bp tandem duplication.

We could not detect any other significant association between the genotype of the 1-FEH IIa 3' duplication and the NRQ values of the four other GH32.

We then investigated a putative quantitative effect of 1-FEH IIa copy number on the carbohydrate-related phenotypic data recorded at harvest and after harvest (Figure 41).

At harvest, no significant association ($p > 0.01$) could be detected between all carbohydrate phenotypic data and any of the NRQ data of the five GH32. As previously reported, at harvest, the 1-FEH IIa 3' duplication did not correlate with any difference of carbohydrate content or properties (Dauchot et al., 2014).

After harvest, both the 1-FEH IIa 3' duplication marker and 1-FEH IIa relative

Loss of function of 1-FEH I**b** has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH I**a**.

copy number, considered at a global level, were strongly and significantly correlated with the variation during the storage of 5 carbohydrate-related parameters (p value ranged from 0.0002 to 10^{-8}). Those two polymorphisms explained an important proportion of the variability of the different carbohydrate-related parameters with R^2 values ranging from 0.36 to 0.60 (Figure 41). The proportion of variance explained by the 1-FEH I**a** 3' duplication (model 1) and 1-FEH I**a** NRQ (model 2) were not significantly different (pval > 0.01) (Supplementary material 8). The results observed for 1-FEH I**a** NRQ are more than likely a reflection of structuration associated with the 1-FEH I**a** 3' duplication, rather than a quantitative effect of 1-FEH I**a** copy number: actually, when the three classes of genotypes (1-FEH I**a** d-/d-, D+/d- and D+/D+) were analyzed independently, no significant association with any carbohydrate-related phenotype was detected within any class. These results suggest that copy number, by itself, has no direct quantitative effect on the susceptibility to post-harvest inulin depolymerization.

7.6.5 Identification of a loss-of-function allele of 1-FEH Ib****

To further characterize the molecular basis of the difference of susceptibility to post-harvest inulin depolymerization, we set out to clone and sequence full length genomic sequences of the 1-FEH I, 1-FEH I**a** and 1-FEH I**b** genes in depolymerization-prone and less susceptible chicory lines. Characterization of the multiple copies of 1-FEH I**a** was of major interest.

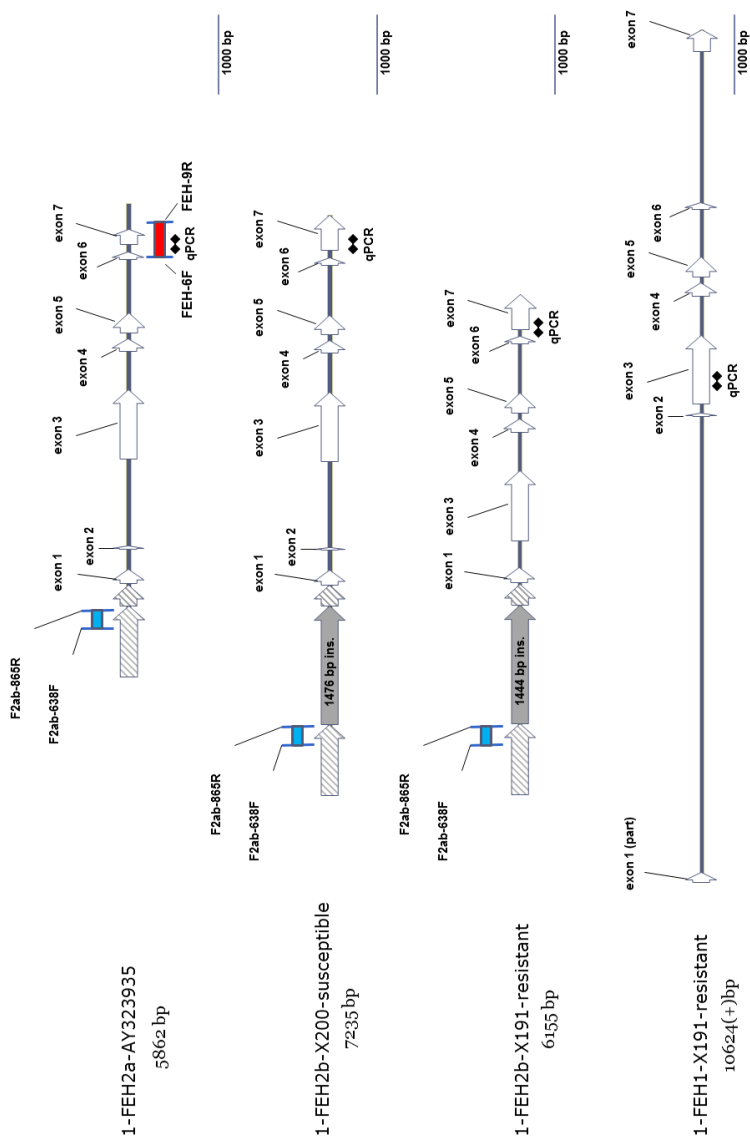
Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Despite several trials with numerous combinations of primers, cycling conditions and the use of long range DNA polymerases, we were unable to re-amplify the previously published 1-FEH IIa genomic sequence (AY323935) in any of our biological samples.

However, using a primer pair initially designed to amplify a large genomic region of 1-FEH IIa (AY323935), we amplified a genomic region covering 100% of the cds of 1-FEH IIb in less susceptible (6154 bp - KM494977) and depolymerization-prone lines (7235 bp - KM494976). Comparison with 1-FEH IIb cDNA sequences (AJ295034) confirmed that these genomic sequences, amplified as single fragments, did correspond to 1-FEH IIb. Previously published partial genomic sequences of 1-FEH IIb (JQ585641 and JQ585640) align with these two larger genomic sequences.

According to the published genomic sequence of 1-FEH IIa (AY323935), 1-FEH IIb genomic sequence shares a number of similarities, including the same intron/exon structure (Figure 42). The only noticeable difference is located in the region upstream of the initiation codon: 1-FEH IIa and 1-FEH IIb present two collinear blocks which are split by a large insertion of approximately 1.5 kb in 1-FEH IIb. When comparing the two 1-FEH IIb alleles (KM494976 and KM494977) isolated from depolymerization prone and resistant lines respectively, we noticed the presence of a large deletion (1018 bp) in KM494977. This deletion includes the 9 bp-long mini-exon 2 and parts of introns 1 and 2. This result is of particular interest since mini-exon 2 contains one of the three conserved active site residues. The loss of this residue must result in the inactivation of the 1-FEH IIb allele (X191 - KM494977).

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH Ila.



Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Figure 42: Gene model for three new genomic sequences of 1-FEH I (KM494975) and 1-FEH IIb (KM494976, KM494977) compared to the previously published gene model of 1-FEH IIa (AY323935). 1-FEH IIb (KM494976) has an intron-exon structure similar to 1-FEH IIa with 6 introns and 7 exons. Promoter region of 1-FEH IIa and 1-FEH IIb share two collinear blocs splitted in 1-FEH IIb by a large insertion of 1.5 kb. Compared to KM494976, the sequence KM494977 presents a 1.028 bp deletion which covers part of introns 1 and 2 and the entire mini-exon 2. Compared to the 1-FEH II group, partial genomic sequence of 1-FEH I is much larger, presents a large first intron and a very short second intron. This structure is typical of cell wall invertases. In 1-FEH II, the second intron is larger than the first one. This structure is more similar to what is observed in vacuolar invertases.

We were also able to amplify a single fragment, a 10,575 bp long region of 1-FEH I in a depolymerization-resistant line (X191 - KM494975) but not in the tested susceptible line (X200). This fragment covers part of the entire cds at the exception of two short regions at the beginning of exon 1 and at the end of exon 7 (Figure 42).

Our cloning strategy, using long range DNA polymerase and unique primer pairs, allowed the amplification and sequencing of new 1-FEH I and 1-FEH IIb gDNA sequences. This strategy ensures that sequences are not chimeric: this point is of particular importance considering the high homology (94%) of the cds of 1-FEH IIa and 1-FEH IIb. The genomic sequence of 1-FEH IIa (AY323935), published much earlier (Michiels et al., 2004), was amplified in three fragments with stringent annealing conditions but poorly specific primer pairs. The overlap between the three fragments was also very limited. The risk

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

that AY323935 is a chimeric sequence, or a rare recombinant allele, is then present and supported by the impossibility to re-amplify this full length sequence as a single fragment in our biological material. Moreover, using primers designed to amplify 1-FEH IIa genomic region (AY323935), we amplified the full length genomic sequence of 1-FEH IIb instead. In the same paper (Michiels et al., 2004), the authors mention that the intron/exon structure of 1-FEH IIa is closer to vacuolar invertase than to cell wall invertase, due to the presence of a large second intron which is typical of vacuolar invertase, while in cell wall invertase, the first intron is larger than the second one. The intron/exon structure of 1-FEH IIb supports its relationship with vacuolar invertase. The gene structure of 1-FEH I, on the contrary, highlights the presence of a very large first intron (5693 bp) and a short second intron (90 bp), which is typical of cell wall invertases (Figure 42). This result further discriminates chicory 1-FEH I from the 1-FEH II a/b group.

7.6.6 Impact of the loss of mini-exon 2 in FEH IIb

We genotyped 112 samples to evaluate the impact of the loss of mini-exon 2 in 1-FEH IIb on the susceptibility to post-harvest inulin depolymerization and to compare these results to those obtained on the same samples with the presence/absence of the 47 bp duplication in the 3'UTR of 1-FEH IIa. The results indicate that these two loci are closely linked. Indeed, except for 8 individuals, the absence of 3' duplication in 1-FEH IIa was associated with the loss of mini-exon 2 in 1-FEH IIb (Supplementary material 9), which results in a very high LD between the two loci ($r^2 = 0.89$).

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

The two loci were both significantly associated with the different carbohydrate characteristics with R^2 values ranging from 0.26 to 0.44 (Supplementary material 10). The proportion of the variance explained by the two polymorphisms was not significantly different for the five carbohydrates parameters ($p_{\text{val}} > 0.01$). Raw genotyping results are presented as supplementary material (Supplementary material 11).

These results indicate that loss of mini-exon 2 in 1-FEH IIb, which results in the loss of one of the three residues of the active site, had a strong impact on the susceptibility to post-harvest inulin depolymerization. This supports the implication of 1-FEH IIb in the depolymerization of inulin induced by exposure to end-season cold temperatures.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

7.7 Discussion

Ultra-high throughput sequencing drastically speeds-up the discovery of SNPs and provides tremendous amounts of new genomic data. All these sequences coupled with the development of CGH arrays highlight the abundance of genomic rearrangements among them the underestimated copy number variants (CNV). Compared to SNP, CNV in the human genome could be at least 1000 up to 10,000 fold more frequent (Lupski, 2007). It is only recently that CNV in animal and plant genomes have been identified as a significant source of phenotypic variation: CNV might represent a new milestone in the understanding of the genetic basis of quantitative traits variability.

However, while we were able to clearly identify copy number variations in the 1-FEH IIa gene, the results presented here provide an example where CNV of functional candidate genes had no quantitative effect on the phenotype.

From a mechanistic point of view, three models are proposed to explain the emergence of CNV in genomes: the Non Allelic Homologous Recombination (NAHR) mostly mediated by low-copy repeats (LCRs), the Non-Homologous End-Joining model (NHEJ) and the Fork Stalling and Template Switching (FoSTeS) model. CNV are further described as recurrent when rearrangements have common size and breakpoints, non-recurrent when CNV share a common region but where breakpoints are located at various positions, and non-recurrent with grouping, when CNV share a common breakpoint which might result from an underlying genomic architecture such as palindrome or

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

cruciform structure (Gu et al., 2008; Stankiewicz and Lupski, 2010).

In our study, we identified variations of copy number only in the 1-FEH IIa allele presenting a 47 bp duplicated region in its 3'UTR (JQ585639). Since we never observed multiple copies of the “short” 1-FEH IIa allele (JQ585638), we hypothesize that this duplication might represent a genomic structure favoring the generation of multiple copies.

Regarding the absence of a quantitative effect of the copy number variation of 1-FEH IIa, we hypothesize that some of the 1-FEH IIa copies might be truncated or pseudogenized and result in non-functional copies. Being located at the 3' end of 1-FEH IIa gene, in case of incomplete copies, the primer pairs we used for CNV estimation cannot discriminate between 5' truncated and full-length functional copies, which prevents us from drawing conclusions on any potential additive effect of additional copies of the gene on the phenotype. To estimate the number of functional 1-FEH IIa copies, one should consider performing qRT-PCR on 1-FEH IIa with sequence-specific primers. However, the very high identity of 1-FEH IIa and 1-FEH IIb cds (96%) makes it very difficult to design 1-FEH IIa specific primer pairs meeting qRT-PCR requirements. In the past, despite extensive trials, we were only able to specifically amplify 1-FEH IIb transcripts for semi-quantitative RT-PCR. At that time, we used very stringent annealing conditions and large amplicons which are incompatible with modern qRT-PCR. Amplification of 1-FEH IIa alone was never achieved (Supplementary material 12).

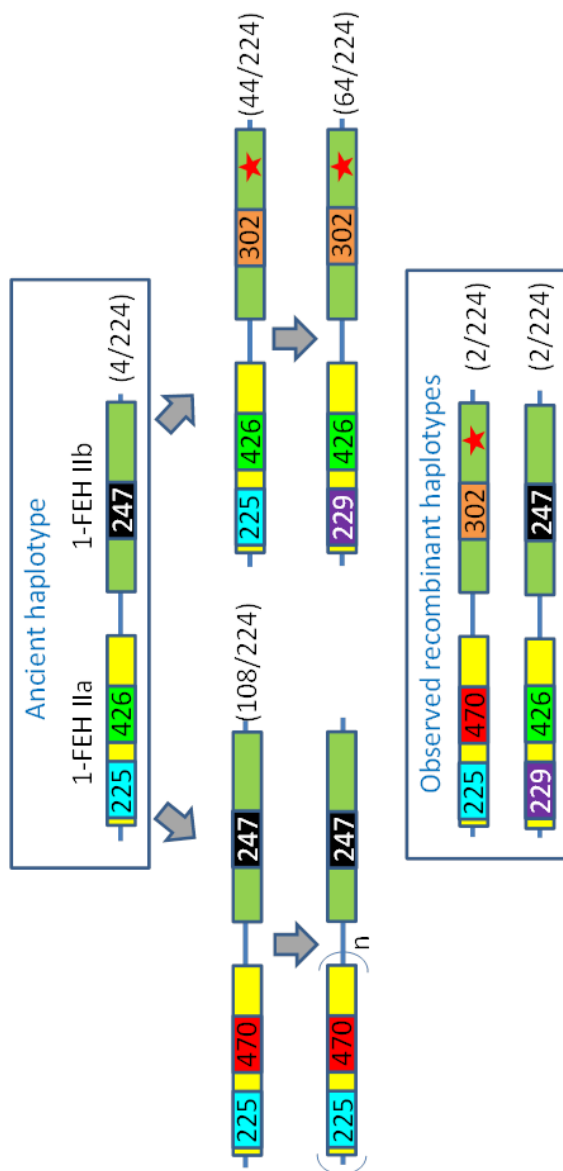
Although copy number variation was not detected in 1-FEH IIb, the sequencing

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

of 1-FEH IIb genomic region revealed the presence of an 1-FEH IIb allele missing the mini-exon 2. This particular allele was observed in chicories less susceptible to post-harvest inulin depolymerization. Genotyping of a larger set of 112 samples revealed that the absence of the duplication in the 3'UTR of 1-FEH IIa was correlated to the absence of multiple copies of 1-FEH IIa and to the loss of mini-exon 2 in 1-FEH IIb. This observation might result from the physical proximity of 1-FEH IIa and 1-FEH IIb, which were previously located on the LG4 of the Rubis 118 map at 28.8 and 30.6 cM respectively (Cadalen et al., 2010).

Based on these results, we identified several haplotypes. We propose an evolutionary scheme to explain the emergence of these haplotypes (Figure 43). Two evolutionary lines are proposed, starting from an ancient haplotype consisting of the “short” 1-FEH IIa allele (426 bp) and a functional 1-FEH IIb allele (identified by 247 bp loci). This ancestral haplotype evolved into a new haplotype harboring a 47 bp tandem duplication in the 3'end of 1-FEH IIa. This haplotype likely evolved and acquired multiple copies of 1-FEH IIa. This 47 bp duplication is considered “recent”, as the tandem repeat is perfectly identical and such duplication was never observed in other Asteraceae (Supplementary material 13). The 1-FEH IIa “copy expansion” is not observed in the second haplotype group (right hand side of Figure 43) which underwent the loss of mini-exon 2 and an insertion in the promoter region (extending it from 247 bp to 302 bp). Later modification of the 1-FEH IIa promoter region (225 bp) which led to a sequence increase up to 229 bp was only observed in this haplotype group. This evolutionary scheme is parsimonious and also explains the existence of two recombinant haplotypes.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.



Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Figure 43: Evolution of the 1-FEH IIa/IIb haplotypes. Numbers reported in the figure refer to the size of PCR products such as scored on an ABI 3130XL fragment analyzer. The haplotypes are defined based on a locus located in the promoter region of 1-FEH IIa (225/229), a locus in the 3'UTR of 1-FEH IIa (426/470) and one locus in the promoter region of 1-FEH IIb (247/302) which was shown to be associated with the loss of mini-exon 2 in the same sequence (red star). The 225/426/247 haplotype is considered ancestral. The 470 allele originates from a tandem duplication of a 47 bp region in the 3'UTR of 1-FEH IIa. This duplication is perfect, which suggests it happened recently. Moreover, this duplication was not detected in other Asteracean 1-FEH II (see Supplementary material 8). This genomic structure might favor the mechanism of CNV since multiple copies of 1-FEH IIa were only observed with this haplotype. The 1-FEH IIb-302 allele was only observed in combination with the ancestral 1-FEH IIa-426 allele. The 1-FEH IIa-229 allele appeared after the loss of mini-exon 2. Four putative recombinant haplotypes were identified. Values under brackets indicate the number of haplotypes detected among 112 diploid individuals.

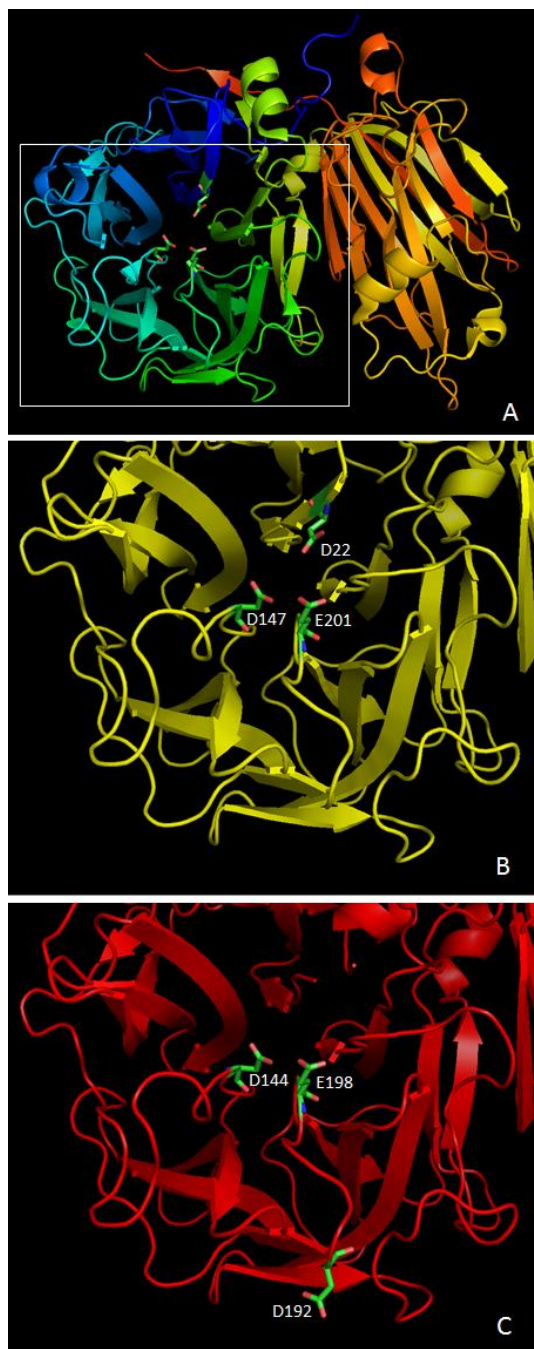
To fully understand the importance of the loss of the 9 bp mini-exon 2 in 1-FEH IIb, one should remember that all GH32 enzymes are composed of a C-terminal β -sheet and a N-terminal β -propeller. In 1-FEH IIa, 3D modeling positions the active site in the center of the β -propeller domain. The active site of 1-FEH IIa is a negatively charged pocket which contains three acidic residues D22 (from the NDPNG conserved sucrose binding box covering mini-exon 2), D147 (from FRDP) and E201 (from WECPD). In the reaction scheme proposed for 1-FEH IIa, E201 protonates the glycosidic oxygen (acid/base catalyst) while the carboxylate group of D22 is used for a nucleophilic attack to produce a covalent fructose enzyme intermediate. The intermediate is then

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

hydrolyzed to release a free fructose, a (n-1) inulin chain and the free enzyme (Verhaest et al., 2005).

The importance of the D residue located in the sucrose binding box was demonstrated as early as 1990 when the mutation of D23 in a yeast Invertase made it basically inactive (Reddy and Maley, 1990). Later, mutation of D23A in AtcwINV1 resulted into a 900-fold reduction of K_{cat} as compared to the wild type enzyme, confirming the crucial role of this residue (Le Roy et al., 2007). In this context, the loss of mini-exon two in 1-FEH IIb, which results into the loss of the DPN triplet inside the sucrose binding box, which itself contains the mandatory D residue, results, more than likely, into an inactive allele of 1-FEH IIb. To further compare both 1-FEH IIb alleles, homology modeling was performed with SWISS-MODEL starting from the available 1-FEH IIa 3D model (1ST8) published earlier (Verhaest et al., 2005). 3D models of both 1-FEH IIb alleles are highly similar (Figure 44a and Figure 44C). The only noticeable differences are the loss of the DPN triplet which results into the loss of the D22 residue in the active site and the apparition of a new D192 residue in the deleted 1-FEH IIb allele. However, this new D residue is located too far from the active site to compensate for the loss of D22. No other mutations on this allele could reasonably restore its activity (Supplementary material 14).

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.



Loss of function of 1-FEH I**b** has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH I**a**.

Figure 44: Homology modeling of the two 1-FEH Ib** alleles identified in this paper was performed using the 3D model of chicory 1-FEH I**a** (1ST8) with SWISS-MODEL online tool. PDB files were manipulated using PyMol. (A) 3D model of the wild type 1-FEH I**b** allele (AIP90173). The active site is located inside the β -propeller structure which contains 3 active residues: D22, D147 and E201. (B) zoom on the active site of AIP90173 (C) model of the active site of the deleted 1-FEH I**b** allele (AIP90174). The loss of the DPN triplet is highlighted by a gap in the 3D model. D144 and E198 have the same positions and orientations as the original D147 and E201 residues. A new D192 residue is observed on the outside of the protein. D192 is too far from the active site to compensate for the loss of D22. AIP90174 is, as a result, more than likely non functional.**

The loss of mini-exon 2 is then of particular interest to explain the lower susceptibility to post-harvest inulin depolymerization. We confirmed experimentally that the loss of 9 bp of the coding sequence of 1-FEH I**b** is not associated with a reading frame shift and that this truncated version can still be transcribed normally (Supplementary material 15).

An increase of the FEH enzymatic activity following exposure to end-season cold temperature has been reported (Van den Ende and Van Laere, 2002). However, at the protein level, the relative contribution of the three FEH enzymes to the total activity is almost impossible to determine. To tackle this question, the same authors published a second paper reporting the seasonal evolution of transcription level of several members of the GH32 multigenic family as analyzed by Northern blot. The results were, at that time, interpreted as illustrative of a higher specificity of cold-regulation of 1-FEH I**a**, as compared to 1-FEH I (Van den Ende et al., 2002). However, results for 1-FEH I**b** were not reported. Considering the high homology between the cds of 1-

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

FEH IIa and 1-FEH IIb (94%), one could reasonably question the specificity of the Northern blot probes used for 1-FEH IIa and consider these Northern blots as illustrative of the expression level of the 1-FEH IIa/b rather than 1-FEH IIa alone. This was taken into account in a later publication where the authors mention that “mRNA blot analysis might also detect other isoforms” (Michiels et al., 2004). In this paper, promoter deletions of 1-FEH IIa and resulting modulation of transcription by cold was studied in detail, but no information regarding the regulation of 1-FEH IIb was mentioned. Personal unpublished results also highlighted the specific transcriptional up-regulation of 1-FEH IIb following exposure to end-season cold temperatures of chicories in the field (Supplementary material 11).

Even if our results do not rule out the specific over-expression of 1-FEH IIa in response to cold-induced inulin depolymerization, they suggest that multiple copies of 1-FEH IIa, as detected by qPCR, have no direct quantitative effect on the susceptibility to post-harvest inulin depolymerization. Questions regarding the non-functional nature of some of these 1-FEH IIa copies remain to be answered. The absence of statistical correlation might result from the presence of truncated or pseudogenized copies of 1-FEH IIa, interfering with the evaluation of a quantitative effect. On the other hand, our results demonstrate that the loss of a functional 1-FEH IIb allele has a direct and significant impact on post-harvest inulin depolymerization and related free sugars content. This underlines the contribution of 1-FEH IIb to inulin depolymerization after harvest. More generally, we suggest that CNVs should be evaluated with several probes located in the promoter and the 5' and 3' regions of the genes of interest to account for possible truncation, eventually coupled with qRT-PCR

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

to provide an accurate estimate of functional copies.

Our results highlight the presence of a non-functional allele of 1-FEH IIb. In this context, association between expression levels and inulin depolymerization has to take into account the nature of the 1-FEH IIb allele. The same remark could apply to 1-FEH IIa if the presence of multiple incomplete copies was confirmed. The methodology and probes developed here can help answering these questions.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

7.8 Conclusions

Here we established for the first time the existence of CNV in the GH32 multigenic family. We also demonstrated that copy number variation, even of a clear candidate gene, is not necessarily responsible for phenotypic variation and needs careful examination of neighboring regions. The discovery of a new loss-of-function allele of 1-FEH IIb whose presence was clearly associated with lower propensity to post-harvest inulin depolymerization supports the implication of 1-FEH IIb in a specific cold stress response adaptation in chicory.

Conflict of interest

O. M. and C. N. are members of Cosucra – Group Warcoing S.A.

Acknowledgements

The authors acknowledge the Walloon Region (D GARNE-Belgium) for supporting this research (Grant D31-1221).

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Ethical standards

The authors acknowledge that the experiments described in this paper comply with the current laws of the country in which they were performed

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

7.9 Reference

- Almal, S.H., and Padh, H. (2011). Implications of gene copy-number variation in health and diseases. *J Hum Genet* 57, 6-13. doi: jhg2011108 [pii] 10.1038/jhg.2011.108.
- Alvarez, C.E., and Akey, J.M. (2011). Copy number variation in the domestic dog. *Mamm Genome* 23, 144-163. doi: 10.1007/s00335-011-9369-8.
- Amaducci, S., and Pritoni, G. (1998). Effect of harvest date and cultivar on *Cichorium intybus* yield components in north Italy. *Ind Crops Prod.* 7, 345-349.
- Cadalen, T., Morchen, M., Blassiau, C., Clabaut, A., Scheer, I., Hilbert, J.L., Hendriks, T., and Quillet, M.C. (2010). Development of SSR markers and construction of a consensus genetic map for chicory (*Cichorium intybus* L.). *Mol Breed.* 25, 699-722. doi: DOI 10.1007/s11032-009-9369-5.
- Clop, A., Vidal, O., and Amills, M. (2012). Copy number variation in the genomes of domestic animals. *Anim Genet* 43, 503-517. doi: 10.1111/j.1365-2052.2012.02317.x.
- D'haene, B., Vandesompele, J., and Hellemans, J. (2010). Accurate and objective copy number profiling using real-time quantitative PCR. *Methods* 50, 262-270. doi: S1046-2023(09)00262-X [pii] 10.1016/j.ymeth.2009.12.007.
- Dauchot, N., Raulier, P., Maudoux, O., Notte, C., Bertin, P., Draye, X., and Van Cutsem, P. (2014). Mutations in chicory FEH genes are statistically associated with enhanced resistance to post-harvest inulin depolymerization. *Theor Appl Genet* 127, 125-135. doi: 10.1007/s00122-013-2206-6.
- De Halleux, S., and Van Cutsem, P. (1997). Cloning and Sequencing of the 1-SST cDNA from chicory Root. *Plant Physiol.* 113, 1003.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

- Debolt, S. (2010). Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol Evol* 2, 441-453. doi: evq033 [pii] 10.1093/gbe/evq033.
- Edelman, J., and Jefford, T.G. (1968). The mechanism of fructosan metabolism in higher plants as exemplified in *Helianthus Tuberosus*. *New Phytol.* 67, 517-531.
- Gibson, G.R., McCartney, A.L., and Rastall, R.A. (2006). Prebiotics and resistance to gastrointestinal infections. *Br J of Nutr* 93, S31-S34.
- Gouesnard, B., Bataillon, T.M., Decoux, G., Rozale, C., Schoen, D.J., and David, J.L. (2001). MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. *J Hered* 92, 93-94.
- Gu, W., Zhang, F., and Lupski, J.R. (2008). Mechanisms for human genomic rearrangements. *Pathogenetics* 1, 4. doi: 10.1186/1755-8417-1-4.
- Helleman, J., Mortier, G., De Paepe, A., Speleman, F., and Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol* 8, R19. doi: gb-2007-8-2-r19 [pii] 10.1186/gb-2007-8-2-r19.
- Hellwege, E.M., Raap, M., Gritscher, D., Willmitzer, L., and Heyer, A.G. (1998). Differences in chain length distribution of inulin from *Cynara scolymus* and *Helianthus tuberosus* are reflected in a transient plant expression system using the respective 1-FFT cDNAs. *FEBS Lett* 427, 25-28.
- Hendry, G.a.F. (1993). Evolutionary Origins and Natural Functions of Fructans - a Climatological, Biogeographic and Mechanistic Appraisal. *New Phytol* 123, 3-14.
- Itaya, N.M., Asega, A.F., Carvalho, M.A., and Figueiredo-Ribeiro Rde, C. (2007). Hydrolase and fructosyltransferase activities implicated in the

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

accumulation of different chain size fructans in three Asteraceae species. *Plant Physiol Biochem* 45, 647-656. doi: 10.1016/j.plaphy.2007.06.004.

Karimi, R., Azizi, M.H., Ghasemlou, M., and Vaziri, M. (2015). Application of inulin in cheese as prebiotic, fat replacer and texturizer: a review. *Carbohydr Polym* 119, 85-100. doi: 10.1016/j.carbpol.2014.11.029.

Keenan, D.F., Resconi, V.C., Kerry, J.P., and Hamill, R.M. (2014). Modelling the influence of inulin as a fat substitute in comminuted meat products on their physico-chemical characteristics and eating quality using a mixture design approach. *Meat Sci* 96, 1384-1394. doi: 10.1016/j.meatsci.2013.11.025.

Knox, A.K., Dhillon, T., Cheng, H., Tondelli, A., Pecchioni, N., and Stockinger, E.J. (2010). CBF gene copy number variation at Frost Resistance-2 is associated with levels of freezing tolerance in temperate-climate cereals. *Theor Appl Genet* 121, 21-35. doi: 10.1007/s00122-010-1288-7.

Le Roy, K., Lammens, W., Verhaest, M., De Coninck, B., Rabijns, A., Van Laere, A., and Van den Ende, W. (2007). Unraveling the difference between invertases and fructan exohydrolases: a single amino acid (Asp-239) substitution transforms *Arabidopsis* cell wall invertase1 into a fructan 1-exohydrolase. *Plant Physiol* 145, 616-625. doi: 10.1104/pp.107.105049.

Lipinski, K.J., Farslow, J.C., Fitzpatrick, K.A., Lynch, M., Katju, V., and Bergthorsson, U. (2011). High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Curr Biol* 21, 306-310. doi: S0960-9822(11)00049-2 [pii] 10.1016/j.cub.2011.01.026.

Lupski, J.R. (2007). Genomic rearrangements and sporadic disease. *Nat Genet* 39, S43-47. doi: 10.1038/ng2084.

Maroufi, A., Van Bockstaele, E., and De Loose, M. (2010). Validation of reference genes for gene expression analysis in chicory (*Cichorium intybus*)

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

using quantitative real-time PCR. *BMC Mol Biol* 11, 15. doi: 1471-2199-11-15 [pii] 10.1186/1471-2199-11-15.

Mendoza, E., Garcia, M.L., Casas, C., and Selgas, M.D. (2001). Inulin as fat substitute in low fat, dry fermented sausages. *Meat Sci* 57, 387-393.

Mensink, M.A., Frijlink, H.W., Van Der Voort Maarschalk, K., and Hinrichs, W.L.J. (2015). Inulin, a flexible oligosaccharide I: Review of its physicochemical characteristics. *Carbohydr Polym.* doi: <http://dx.doi.org/10.1016/j.carbpol.2015.05.026>.

Meyer, D., and Stasse-Wolthuis, M. (2009). The bifidogenic effect of inulin and oligofructose and its consequences for gut health. *Eur J Clin Nutr* 63, 1277-1289. doi: 10.1038/ejcn.2009.64.

Michiels, A., Van Laere, A., Van den Ende, W., and Tucker, M. (2004). Expression analysis of a chicory fructan 1-exohydrolase gene reveals complex regulation by cold. *J Exp Bot* 55, 1325-1333. doi: 10.1093/jxb/erh153.

Monti, A., Amaducci, M.T., Pritoni, G., and Venturi, G. (2005). Growth, fructan yield, and quality of chicory (*Cichorium intybus* L.) as related to photosynthetic capacity, harvest time, and water regime. *J. Exp Bot* 56, 1389-1395.

Nicholas, T.J., Baker, C., Eichler, E.E., and Akey, J.M. (2011). A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. *BMC Genomics* 12, 414. doi: 1471-2164-12-414 [pii] 10.1186/1471-2164-12-414.

R Development Core Team (2013). "R: A language and environment for statistical computing." (Vienna, Austria: R Foundation for Statistical Computing).

Reddy, V.A., and Maley, F. (1990). Identification of an active-site residue in yeast invertase by affinity labeling and site-directed mutagenesis. *J Biol Chem* 265, 10817-10820.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Springer, N.M., Ying, K., Fu, Y., Ji, T., Yeh, C.T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A.L., Barbazuk, W.B., Jeddeloh, J.A., Nettleton, D., and Schnable, P.S. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet* 5, e1000734. doi: 10.1371/journal.pgen.1000734.

Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med* 61, 437-455. doi: 10.1146/annurev-med-100708-204735.

Van Arkel, J., Vergauwen, R., Sevenier, R., Hakkert, J.C., Van Laere, A., Bouwmeester, H.J., Koops, A.J., and Van Der Meer, I.M. (2012). Sink filling, inulin metabolizing enzymes and carbohydrate status in field grown chicory (*Cichorium intybus* L.). *J Plant Physiol* 169, 1520-1529. doi: 10.1016/j.jplph.2012.06.005.

Van den Ende, W., Michiels, A., De Roover, J., and Van Laere, A. (2002). Fructan biosynthetic and breakdown enzymes in dicots evolved from different invertases. Expression of fructan genes throughout chicory development. *ScientificWorldJournal* 2, 1281-1295. doi: 10.1100/tsw.2002.288.

Van den Ende, W., Michiels, A., De Roover, J., Verhaert, P., and Van Laere, A. (2000). Cloning and functional analysis of chicory root fructan1-exohydrolase I (1-FEH I): a vacuolar enzyme derived from a cell-wall invertase ancestor? Mass fingerprint of the 1-FEH I enzyme. *Plant J* 24, 447-456.

Van den Ende, W., Michiels, A., Van Wouterghem, D., Clerens, S.P., De Roover, J., and Van Laere, A.J. (2001). Defoliation induces fructan 1-exohydrolase II in Witloof chicory roots. Cloning and purification of two isoforms, fructan 1-exohydrolase IIa and fructan 1-exohydrolase IIb. Mass fingerprint of the fructan 1-exohydrolase II enzymes. *Plant Physiol* 126, 1186-1195.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Van den Ende, W., and Van Laere, A. (2002). Induction of 1-FEH in mature chicory roots appears to be related to low temperatures rather than to leaf damage. *ScientificWorldJournal* 2, 1750-1761. doi: 10.1100/tsw.2002.857.

Van Laere, A., and Van den Ende, W. (2002). Inulin metabolism in dicots: chicory as a model system. *Plant Cell Environ* 25, 803-813.

Van Waes, C., Baert, J., Carlier, L., and Van Bockstaele, E. (1998). A rapid determination of the total sugar content and the average inulin chain length in roots of chicory (*Cichorium intybus* L). *J Sci Food Agric* 76, 107-110.

Verhaest, M., Van den Ende, W., Le Roy, K., De Ranter, C.J., Van Laere, A., and Rabijns, A. (2005). X-ray diffraction structure of a plant glycosyl hydrolase family 32 protein: fructan 1-exohydrolase IIa of *Cichorium intybus*. *Plant J* 41, 400-411. doi: DOI 10.1111/j.1365-313X.2004.02304.x.

Wilson, R.G., Smith, J.A., and Yonts, C.D. (2004). Chicory root yield and carbohydrate composition is influenced by cultivar selection, planting, and harvest date. *Crop Sci.* 44, 748-752.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

7.10 Supplementary materials

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

	DP.R ²	DP.Pval	IN.R ²	IN.Pval	Suc.R ²	Suc.Pval	Fru.R ²	Fru.Pval	Glu.R ²	Glu.Pval
indel FEH IIa	0,5028	3,72E-06	0,5754	2,75E-07	0,5969	1,17E-07	0,3567	0,0002614	0,4011	8,03E-05
NRQ FEH IIa	0,5246	3,58E-07	0,5454	1,65E-07	0,5366	2,30E-07	0,377	4,07E-05	0,4425	5,77E-06

Supplementary material 8: Comparison of the variance explained by the

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

presence/absence of a 47 bp in the 3'UTR of 1-FEH IIa and the normalized relative copy number of the same gene (as estimated by qPCR). Depending on the carbohydrate parameters (DP – degree of polymerization, IN – inulin content, Suc – free sucrose content, Fru – free fructose content, Glu – free glucose content), the two models explained from 36% up to 60% of the variance. 1-FEH IIa copy number did not increase the proportion of the variance explained by the indel model.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

		1-FEH IIa		
		d-/d-	D+/d-	D+/D+
1-FEH IIb	d-/d-	0	6	36
	D+/d-	0	31	1
	D+/D+	37	1	0

Supplementary material 9: Correlation table obtained for 112 individual for two genetic markers located in 1-FEH IIa and 1-FEH IIb. The two markers are highly correlated ($r^2 = 0.89$). “d-/d- 1-FEH IIa” stands for “homozygous for the absence of duplication” in the 3' UTR of 1-FEH IIa, “D+/D+ 1-FEH IIa” stands for “homozygous for the presence of duplication” in the 3' UTR of 1-FEH IIa and “d-/D+” corresponds to heterozygous genotypes. “D+/D+ 1-FEH IIb” stands for “homozygous for the loss of mini-exon 2 in 1-FEH IIb”, “d-/d-” stands for homozygous for the normal 1-FEH IIb allele” and “d-/D+” corresponds to heterozygous genotypes.

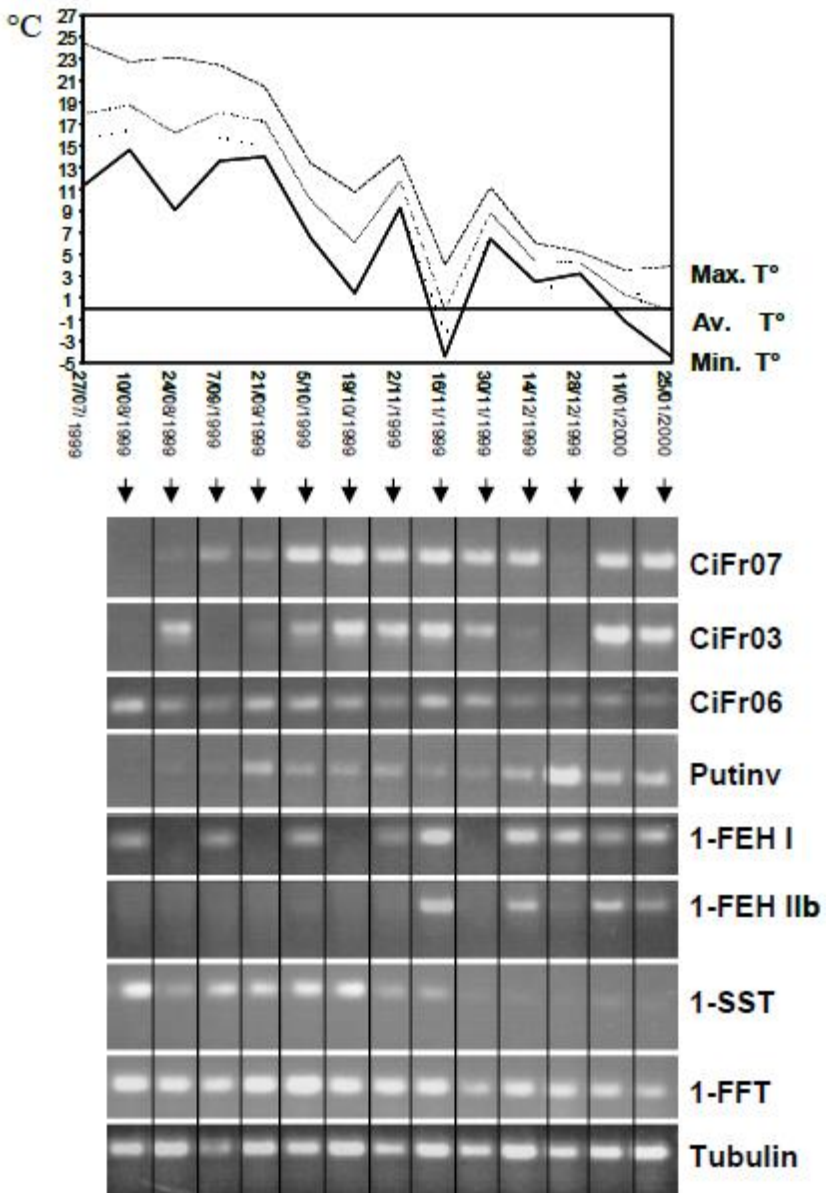
Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Target	DP		IN		SUCn		FFn		FGn	
	R ²	Pval	R ²	Pval	R ²	Pval	R ²	Pval	R ²	Pval
Indel 1FEH IIa	0,30	5,3E-09	0,42	8,2E-14	0,43	4,0E-14	0,31	2,2E-09	0,31	2,2E-09
Ex2 1FEH IIb	0,26	6,3E-08	0,44	2,4E-14	0,43	3,2E-14	0,32	5,3E-10	0,31	1,7E-09

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Supplementary material 10: For five parameters related with carbohydrate metabolism, proportion of the variance explained by an indel located in the 3'UTR of 1-FEH IIa and by the loss of mini-exon 2 in 1-FEH IIb. P values are not significant which indicates a strong correlation between both polymorphisms. Genotyping results obtained for 3 loci located in 1-FEH IIa and 1-FEH IIb on 112 individuals. Amplification products were scored on a ABI3130 XL fragment analyzer with 500-LIZ size standard. For each diploid locus, Supplementary material 15 presents the identified alleles that were scored. The first two columns are presenting the genotyping results of the 112 samples for the presence or absence of duplication in the 3' UTR of 1-FEH IIa. Absence of duplication is scored by a 427 bp long amplicon, while the presence of the duplication generates a 470 bp long fragment. Columns to 6 corresponds to the peaks detected after analyzing the amplification products obtained with a single primer pair amplifying simultaneously part or the promotor of both 1-FEH IIa and 1-FEH Ib. Fragments 225 and 229 are amplicons resulting from the amplification of part of 1-FEH IIa promotor, while fragments of 247 and 302 bp long result from the amplification of 1-FEH IIb promoters. These results were used to propose the evolution of FEH II haplotypes.

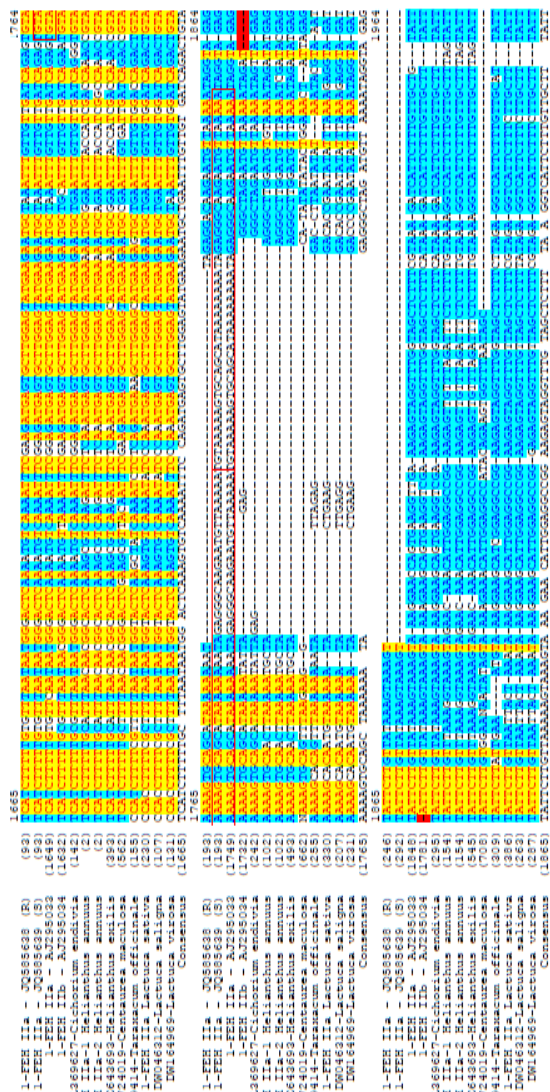
Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.



Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Supplementary material 11: Semi-quantitative RT-PCR analyses of fructan synthesis and degradation enzymes transcribed in chicory roots during growing season 1999-2000. Expression profiles correspond to 13 sampling dates for which respective minimal, average and maximal temperatures are presented. All the root tissues were sampled in field. RNA was normalized to a concentration of 1µg/µl. Tubulin was used as standard. As observed, 1-FFT transcription was constant during growing season. 1-SST expression started decreasing after exposure to temperatures lower than 4°C while, at the same time, 1-FEH IIb was overexpressed. 1-FEH IIb seemed to be expressed more specifically in response to exposure to cold temperatures, while 1-FEH I transcripts could be detected throughout growing season

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.



Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Supplementary material 12: Alignment of 1-FEH IIa putative orthologs retrieved after comparing AJ295033 against the NCBI EST database (BLASTn). 1-FEH IIa-1 *Helianthus annuus* is the contig of BU035623, GE492441, GE495456 and BU034394. 1-FEH IIa-2 *Helianthus annuus* is the contig of BU035062, BU031314, DY930605, DY921879, DY922023, DY955464 and BY03132. 1-FEH IIa *Lactuca sativa* is the contig of DW137787 and BQ844426. 1-FEH IIa tandem repeats are indicated by the two red boxes. This alignment indicates that the 47 bp duplication is absent from all publicly available 1-FEH II putative orthologs.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Supplementary material 13:Alignment of the amino-acid sequence of the 3 known chicory FEH + the two newly isolated FEH (this paper), including the « wild type » long FEH-IIb and the mutated (short) FEH IIb allele, missing the mini-exon 2 (DPN residues). Positions of the 3 residues from the active site are highlighted by black boxes. Other mutations in the “short” FEH IIb allele are unlikely to restore the activity of this FEH allele.

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Supplementary material 14: At 55°C, primer pair F2b-82Fw and F2b-231Rv amplified both 150 bp long 1-FEH IIa cDNA and 1-FEH IIb allele missing the mini-exon 2 (scored as a 140 bp long fragment). cDNA was extracted from a chicory root collected in January 2012 after exposure to winter cold temperatures in a genotype homozygous for the 1-FEH IIb allele missing the mini-exon 2. This deleted allele is normally transcribed in the root. 1-FEH I was not amplified (expected 146 bp). Peak around 180 bp is an artefact due to the use of labelled dUTP during the PCR.

F2b-82Fw CGGGATCTGAATGATGTGATACTA (1-FEH IIb, AJ295034, 85-105) left black box

F2b-231Rv GTATTGGTAGAAGAAATGGTACACTCCT (1-FEH IIb, AJ295034, 204-231) right black box

Mini-exon ATCCCAATG (1-FEH IIb, AJ295034, 179-187) red box

Expected amplification size 150 bp (full – 1-FEH IIb and 1-FEH IIa), 141 bp (1-FEH IIb without 9 bp mini-exon 2). PCR with FAM labelled

dUTP.

Analyzed on ABI 3130XL device with 500-Liz size standard and POP-4 polymer. Separation at 60°C.

Loss of function of 1-FEH IIB has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIA.

	F2a-indel		F2a-prom		F2b-prom	
	426	470	225	229	247	302
X073	426,47	426,47	225,34			302,41
X070	426,4	426,04	225,35			302,47
X091	426,54	426,54	225,38			302,49
X040	426,62	426,62	225,49			302,59
X040	426,62	426,62	225,49			302,59
X014	426,53	426,53	225,51			302,62
X191	426,59	426,59	225,51			302,64
X080	426,48	426,48	225,53			302,47
X018	426,36	426,36	225,55			302,66
X013	426,39	426,39	225,59			302,56
X172	426,44	426,44	225,68			302,77
X164	426,7	426,7	225,67	229,42		302,77
X195	426,5	426,5	225,53	229,35		302,55
X126	426,25	426,25	225,45	229,3		302,46
X190	426,54	426,54	225,46	229,28		302,54
X076	426,52	426,52	225,5	229,28		302,58
X100	426,52	426,52	225,47	229,24		302,48
X083	426,51	426,51	225,48	229,24		302,48
X001	426,45	426,45	225,44	229,18		302,85
X137	426,58	426,58	225,38	229,16		302,4
X140	426,48	426,48	225,37	229,14		302,48
X128	426,53	426,53	225,38	229,09		302,4
X020	426,04	426,04		229,28		302,56
X102	426,35	426,35		229,26		302,56
X057	426,41	426,41		229,23		302,57

Loss of function of 1-FEH IIB has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIA.

X066	426,41	426,41		229,04		302,49
X197	426,43	426,43		229,34		302,63
X129	426,47	426,47		229,15		302,49
X104	426,48	426,48		229,19		302,05
X104	426,48	426,48		229,19		302,05
X193	426,49	426,49		229,36		302,62
X061	426,5	426,5		229,26		302,58
X094	426,51	426,51		229,26		302,47
X124	426,52	426,52		229,13		302,5
X136	426,52	426,52		229,08		302,39
X051	426,57	426,57		229,02		302,57
X101	426,57	426,57		229,15		302,58
X135	426,58	426,58		229,15		302,49
X194	426,58	426,58		229,24		302,62
X067	426,52	470,93	225,39	229,08	247,62	
X112	426,63	470,85	225,53	229,11	247,7	302,39
X056	426,63	470,89	225,43	229,12	247,72	302,49
X123	426,47	470,84	225,44	229,13	247,65	302,49
X133	426,46	470,87	225,45	229,15	247,72	302,58
X077	426,55	470,93	225,36	229,16	247,65	302,54
X077	426,55	470,93	225,36	229,16	247,65	302,54
X077	426,55	470,93	225,36	229,16	247,65	302,54
X114	426,54	470,94	225,38	229,16	247,61	302,5
X007	426,46	470,71	225,46	229,17	247,86	302,76
X107	426,52	470,87	225,41	229,17	247,67	302,48
X030	426,51	470,89	225,39	229,18	247,71	302,49
X032	426,53	471,07	225,39	229,18	247,63	302,4
X132	426,58	470,9	225,45	229,21	247,75	302,48
X147	426,55	471,02	225,47	229,21	247,71	302,46

Loss of function of 1-FEH I1b has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH I1a.

X156	426,46	470,74	225,46	229,23	247,76	302,48
X131	426,56	470,8	225,47	229,24	247,7	302,48
X098	426,52	470,8	225,45	229,25	247,75	302,51
X138	426,62	470,87	225,49	229,25	247,73	302,47
X103	426,56	470,96	225,49	229,25	248,7	302,47
X064	426,6	470,9	225,53	229,26	247,76	302,55
X187	426,54	470,71	225,74		248,08	302,83
X169	426,44	470,77	225,68		247,92	302,76
X145	426,52	470,78	225,39		247,75	302,49
X012	426,61	470,8	225,62		247,91	302,64
X165	426,52	470,81	225,61		247,87	302,7
X078	426,5	470,96	225,38		247,71	302,47
X127	426,51	470,83	225,45		247,73	302,48
X027	426,53	470,85	225,46		247,73	302,48
X006	426,53	470,88	225,52		247,8	302,56
X081	426,45	470,87	225,4		247,76	302,57
X163	426,55	470,94	225,66		248,03	302,78
X154	426,53	470,81	225,45		247,74	
X186	426,49	471	225,65		248,05	
X158	426,68	471,01	225,44		247,74	
X189	426,59	471,08	225,56		247,79	
X120	470,55	470,55	225,38		247,7	
X173	470,67	470,67	225,65		247,93	
X003	470,69	470,69	225,52		247,89	
X065	470,74	470,74	225,21		247,74	
X087	470,78	470,78	225,43		247,68	
X087	470,78	470,78	225,43		247,68	
X087	470,78	470,78	225,43		247,68	
X118	470,78	470,78	225,41		247,76	

Loss of function of 1-FEH IIB has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIA.

X174	470,78	470,78	225,7		247,95	
X179	470,78	470,78	225,57		247,89	
X046	470,8	470,8	225,45		247,74	302,55
X109	470,8	470,8	225,44		247,74	
X075	470,81	470,81	225,38		247,72	
X161	470,81	470,81	225,45		247,77	
X021	470,83	470,83	225,5		247,77	
X026	470,83	470,83	225,63		247,89	
X060	470,83	470,83	225,49		247,72	
X111	470,83	470,83	225,51		247,75	
X144	470,83	470,83	225,52		247,79	
X079	470,84	470,84	225,37		247,77	
X180	470,84	470,84	225,64		248	
X141	470,85	470,85	225,45		247,75	
X183	470,85	470,85	225,5		247,85	
X009	470,86	470,86	225,34		248,52	
X152	470,87	470,87	225,42		247,78	
X024	470,88	470,88	225,46		247,8	
X072	470,89	470,89	225,42		247,69	
X175	470,89	470,89	225,7		248,07	
X041	470,91	470,91	225,48		247,82	
X119	470,91	470,91	225,49		247,74	
X177	470,91	470,91	225,64		248	
X082	470,94	470,94	225,46		247,76	
X015	470,96	470,96	225,55		247,77	
X181	470,96	470,96	225,55		247,86	
X036	470,97	470,97	225,42		247,68	
X034	470,99	470,99	225,46		247,76	
X162	470,99	470,99	225,43		247,76	

Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa.

Supplementary material 15: Genotyping results obtained for 3 loci located in 1-FEH IIa and 1-FEH IIb on 112 individuals. Amplification products were scored on a ABI3130 XL fragment analyzer with 500-LIZ size standard. For each diploid locus, table presents the identified alleles that were scored. The first two columns are presenting the genotyping results of the 112 samples for the presence or absence of duplication in the 3' UTR of 1-FEH IIa. Absence of duplication is scored by a 427 bp long amplicon, while the presence of the duplication generates a 470 bp long fragment. Columns 3 to 6 corresponds to the peaks detected after analyzing the amplification products obtained with a single primer pair amplifying simultaneously part of the promoter of both 1-FEH IIa and 1-FEH IIb. Fragments 225 and 229 are amplicons resulting from the amplification of part of 1-FEH IIa promoter, while fragments of 247 and 302 bp long result from the amplification of 1-FEH IIb promoters. These results were used to propose the evolution of FEH II haplotypes (Figure 43).

8 GWAS analysis of carbohydrate and inulin quality in chicory

P. Raulier, N. Dauchot, O. Maudoux, C. Notté, P. Van Cutsem, P. Bertin and
X. Draye

8.2 Abstract

Chicory is the main industrial source of inulin. The recent identification of an indel in *1-FEH IIb* accounting for up to 35 % of the phenotypic variance of the susceptibility to cold-induced inulin degradation has been a major step in deciphering the genetic determinism of inulin and carbohydrates content in chicory. The present study focuses on the genetic determinism of inulin and carbohydrates content at harvest and of the sensitivity to cold-induced inulin degradation. A genome-wide association study (GWAS) was performed on a diversity panel of 366 lines representing the germplasm from which current cultivars originate. A total of 122 650 SNPs were identified by Genotyping-by-Sequencing (GBS), of which 9 859 SNPs presented a sufficient minor-allele frequency and calling rate to integrate the GWAS. These SNPs were mapped on a genome draft of 21 680 contigs covering together 1.1 GB of the 1.3 GB of the chicory genome. The diversity panel was characterized by a very low linkage disequilibrium (LD) extent estimated on the genome scale ($r^2 < 0.2$ at 300 bp). The LD extended on long distance in the region surrounding *1-FEH IIb* ($r^2 = 0.48$ at 360 Kb), reflecting a hitchhiking event. On the synthesis side, the GWAS identified at least 6 SNPs significantly associated to carbohydrates and inulin quality at harvest. On the degradation side, several SNPs physically close to the indel in *1-FEH IIb* and one independent SNP were significantly associated to the cold-induced degradation.

8.3 Introduction

Industrial chicory is the main commercial source of inulin, a linear fructose polymer used by the agroindustry as texturizer, fat substitute and low calories sweetener (Mendoza et al., 2001; Keenan et al., 2014; Karimi et al., 2015). Inulin is also a dietary fiber used for its health-promoting properties (Gibson et al., 2006; Meyer and Stasse-Wolthuis, 2009). The value and industrial properties of the chicory root extract are defined by the mean number of fructosyl units making up inulin molecules (mDP, mean polymerization degree) and by the concentration in free fructose. The industry is interested in high mDP inulin (Van Arkel et al. 2013) and in low fructose content which cannot be sold on the European market since the revision of the European sugar regime in 2006. Moreover, fructose hampers the process of inulin purification.

The mDP and free fructose content result from the equilibrium between inulin synthesis and degradation processes. Van Arkel (2012) described three phases in the inulin metabolism. The first phase begins one month after sowing with the radial growth of the roots and involves the activity of two enzymes. On the one side, the sucrose:sucrose 1-fructosyltransferase (1-SST) mediates the transfer of a fructosyl unit from a sucrose (F-G) to another sucrose, releasing the trisaccharide fructan 1-kestose (G-F-F) and a single glucose. On the other side, the fructan:fructan 1-fructosyltransferase (1-FFT) transfers fructosyl units from a 1-kestose to another fructan. The joint activities of 1-SST and 1-FFT result in the synthesis of high mDP inulin. The

second phase starts around mid-September (in the Netherlands). During this phase, the 1-SST activity declines while the back-transfer activity of 1-FFT increases. The latter is the transfer of a fructosyl unit from a fructan to a sucrose or to a shorter fructan, resulting in the accumulation of shorter inulin molecules. The third phase starts with the low autumnal temperatures that trigger the fructan exohydrolase (1-FEH) activity. This occurs around mid-November in the Netherlands. The 1-FEH hydrolyzes inulin, resulting in a decrease in mDP and the release of the undesirable free fructoses. In Belgium, France and the Netherlands, roots are generally harvested from September to December, resulting in loss in inulin quality.

The genetic bases of the variability of fructan content and/or quality have been investigated in wheat, barley, ryegrass and onion. In wheat, two major-effect QTLs included genes encoding for enzymes of the fructan metabolism (Huynh et al. 2008, 2012; Zhang et al. 2008a): one major QTL corresponded to the location of a cluster including *1-SST*, *1-FFT* and *6-SFT*, all under the control of a unique transcriptional activator (TaMYB13); the other major QTL corresponded to the *6-FEH* location. A few minor-effect QTLs were also detected but could not be associated to any candidate gene. In onion, one major QTL included a gene encoding for a sucrose-phosphate synthase that may influence the sucrose level. Sucrose level may affect 1-SST expression, as observed in others species (Müller et al. 2000; Wang, Van den Ende, and Tillberg 2000). In ryegrass, a candidate gene-based association study revealed an association between fructan content and a polymorphism in a gene functionally associated with stem elongation (*LpCAD2*). The underlying mechanism is still unknown but the authors suggest that this gene may affect

carbon partitioning in the plant (Pembleton et al. 2013).

In chicory, inulin yield and mDP at harvest and the sensitivity to cold induced inulin degradation are cultivar-dependent (Baert et al. 1997, Amaducci et al. 1998, Wilson et al. 2004). In an effort to understand the genetic basis of the sensitivity to cold-induced inulin degradation, Dauchot et al. (2014 and 2015) investigated polymorphisms within nine fructosyl transferases from the family 32 glycoside hydrolase (GH32, including *I-SST*, *I-FFT* and *I-FEH*). Among them, the presence/absence of a 9 bp mini exon in an active site of *I-FEH IIb* was highly associated with the sensitivity to cold-induced inulin degradation and accounted for up to 35% and 18 % of the variance in variation of inulin content and mDP after cold exposure, respectively (Dauchot et al. 2014). No polymorphism in the other GH32 genes could be associated with variables describing inulin quality at harvest or following degradation. Other investigations showed that the expression level of 1-FEH varied across individuals and cultivars, however that expression level was not correlated with carbohydrate content (Maroufi et al. 2010 and Maroufi et al. 2012). To our knowledge, the understanding of the genetic determinism of fructans and carbohydrates is limited to one 9 bp indel within 1-FEH IIb highly associated to the sensitivity to the cold-induced inulin degradation and the absence of association between polymorphisms within 8 other genes of the GH32 family and any trait describing the fructan and carbohydrate quality (Dauchot et al 2014 and 2015). Investigation in onion and ryegrass showed that fructan determinism may involve major and minor QTLs that do not correspond to fructan genes. Therefore, in chicory, a genome-wide investigation may contribute to identify novel QTLs and uncover pleiotropic

effects.

The aim of the present study was to analyze the genetic determinism of the yield and quality of carbohydrates and inulin at harvest and cold-induced inulin degradation. In this context, we implemented a genome-wide association mapping (GWAS) on a panel of 366 individuals selected out of the collection of industrial chicory described in Raulier et al. (2015). The large number of markers required was achieved using Genotyping-by-Sequencing technologies (Elshire et al. 2012). To anchor these markers on a genetic map, we initiated the sequencing of the chicory genome. Unfortunately, the complete assembly of the genome is not available yet and informations relative to the SNPs position are incomplete.

8.4 Materials and methods

8.4.1 Plant material

The study was performed on a diversity panel of 366 industrial chicory individuals. These individuals were sampled out of a collection of 1297 individuals from 15 cultivars described in Raulier et al. (2015) (Table 14). Those cultivars constitute the origin of modern industrial chicory cultivars. The selection was achieved with the MSTRAT algorithm (Gouesnard et al. 2001) based on the 15 SSRs markers used in Raulier et al. (2015) and in Dauchot et al. (2014). The MSTRAT algorithm is a sampling method based on molecular markers used to maximize the allelic richness and genetic diversity of a group of sampled individuals. The SSR genotyping was performed following Muys et al. (2013).

Cultivar	Ni collection	Ni panel	Breeding institution	Country	Year of release
'Pévèle'	81	21	S.A. Florimond Desprez (*)	France	1984
'Fruitosa'	78	17	S.A. Florimond Desprez (*)	France	1984
'Cassel'	77	23	S.A. Florimond Desprez (*)	France	1986
'Orchies'	84	24	S.A. Florimond Desprez (*)	France	1987
'Tilda'	75	26	Bataille Alphonse (*)	France	1981
'Fredonia'	71	23	Sarea GmbH (*)	Austria	
'Polanowicka'	123	29	Centrum Obrotu Rolnego "Agro-Kal" sp. z o.o. (*)	Poland	
'Kujawska'	103	42	Gospodarstwo Rolno- Nasienne KOW-MAR Maria J. Kowalska (*)	Poland	
'Dageraad'	116	25	Merelbeke (*)	Belgium	1981
'Halle'	120	29	Merelbeke (*)	Belgium	
'Sabau 3'	75	20	Ets Sabau (*)	France	
'Bergues'	72	25	S.A. Florimond Desprez (*)	France	1988
'Candi'	75	16	Selbat (*)	France	
'Novipa'	73	40	Merelbeke (*)	Belgium	1956
'Hicor'	74	13	Bataille Alphonse (*)	France	

Table 14: Constitution of the diversity panel assembled for the study: number of individuals in the collection described in Raulier et al. (2015) (Ni collection), number of individuals selected in the panel used in this study (Ni panel), the breeding institution, the country of origin and the year of release of the cultivar.

8.4.2 Phenotyping

The phenotyping was performed on 366 lines stemming from the selfing of the 366 individuals of the diversity panel. Each of the 366 lines was phenotyped during one of three field trials realized between 2011 and 2013. The first trial included 112 lines from all 15 cultivars, the second and third trials included the remaining lines from 9 and 6 cultivars, respectively (Table 15). The lines phenotyped in the first trial were the same as in Dauchot et al. 2013 and Dauchot et al. 2014. The experimental plots consisted in one central row including the line of interest flanked by two rows of a commercial cultivar (Dolce). The rows were 5.2 m long and spaced 45 cm from each other. Within the row, plants were spaced 13 cm from each other. A maximum of 24 roots were harvested per experimental plot. The trials were sown between the end of April and the beginning of May, in the neighborhood of Warcoing (7740, Belgium, 50° 42' North 3° 20' East) (Table 15). They were harvested in the beginning of October. Each trial was laid out as a randomized split-plot design.

	Trial 1	Trial 2	Trial 3
Sowing date	2/05/2011	14/05/2012	18/04/2013
Harvest	From 2/10/2011 to 9/10/2011	From 2 /10/2012 to the 9/10/2012	From the 23/09/2013 to the 3/10/2013
Location	Pecq (7740, Belgium)	Hérinnes (7742, Belgium)	Hérinnes (7742, Belgium)
Soil quality	Loam	Sandy loam	Sandy loam
Number of lines	112	122	139
Cultivars at the origin of lines	Pévèle, Fruitosa, Cassel, Orchies, Sabau 3, Bergues, Candi, Novipa, Hicor, Tilda, Fredonia, Polanowicka, Kujawska, Dageraad and Halle	Pévèle, Fruitosa, Cassel, Orchies, Sabau 3, Bergues, Candi, Novipa and Hicor.	Tilda, Fredonia, Polanowicka, Kujawska, Dageraad and Halle

Table 15: Sowing date, harvest date, location, soil quality, number of line, and cultivar at the origin of the lines implanted in the three trials.

Due to the destructive nature of the carbohydrate phenotyping, each line was split in two groups of roots. The carbohydrate characterization of the first group was performed at harvest and that of the second group was performed after an exposure to cold, to simulate low temperatures occurring in the field in autumn, as described in Dauchot et al. (2014). The roots from each group were pooled and analyzed together as one sample. The carbohydrate characterization was performed by HPLC according to Van Waes (Van Waes et al., 1998) and yielded the following variables: inulin degree of

polymerization (DP_{in}), inulin content (IN), sucrose content (SUC_n), free fructose content (FF_n) and free glucose content (FG_n). The IN, SUC_n, FF_n and FG_n were expressed as a percentage of the total carbohydrate mass.

The fresh weight (FW), the dry matter content (DMC) and the carbohydrate content (CC measured in percentage of DMC) were determined as described in Dauchot et al. (2014).

Except for the FW, we present the result of variables measured at harvest and the variation of these variables after the cold exposure computed as (value measured at harvest) – (value measured after cold exposure).

In presence of inulo-*n*-ose, such as observed after cold-induced degradation result in an over estimation of the DP. In order to circumvent this limitation, the DP in the second and third trial were estimated by high-performance anion exchange chromatography coupled with pulse amperometric detection (HPAEC-PAD) on a Dionex DX500 chromatographic system chromatographic. This procedure analyse the generate complete fructan profile (inulin and inulon-ose) and unable to generate accurate DP. The HPAEC-PAD was realized according Blecker et al.2002.

8.4.3 Phenotypic analysis

A mixed model analysis was performed prior to the GWAS analysis in order to obtain phenotypic values corrected for the trial effect. Trial was modelled

as a fixed factor and cultivar as a random factor. The analysis was performed with the lmer function from the R package lme4 (Bates et al. 2014). Most variables (all except the variation of DP_{in} after cold exposure) did not show any significant trial $\times$ cultivar interaction. Therefore, for variables presenting a significant trial effect, the phenotypic value of the second and third trial were adjusted to the mean of the first trial. Uncorrected values were used for variables with no significant trial effect.

8.4.4 Genome sequencing

On homozygous individual stemming from 8 selfing cycle was sequenced by a combination Illumina shorter reads (30 Gbases of 2x100nt Illumina HiSeq 2000, 15 Gbases 2x300nt of Miseq, 70 Gbases of 2x150nt Illumina NExtSeq) and 26 Gbases PACBio RSII P6C4 long reads. In order to assemble the chicory reads we performed a QC for each fastq file using the 'FASTQC' software. Depending on the outcome adapter removal and quality, trimming where run on each pair of reads. Typically a quality of Fred-score = 20 was used as minimum threshold. The resulting reads were then used using the CLCBio assembly cell (Kmer 58, bubble 50). This software is a Debruyrn-graph-approach based assembler good for the massive amount of short reads obtained from Illumina. Although the details of the algorithm are not disclosed, the advantage of that software is that it produces relatively good contigs, while using moderate hardware resources (less RAM and time compared to many open source assembly software). The output of the CLCBio assembler is a set of contigs (multi-fasta file). The PACBio reads

where "cleaned" using the new software LoRDEC, that uses the Illumina short reads to correct the longer reads, known to be much less accurate. The contigs and cleaned PACBio reads were combined in a hybrid assembly using DBG2OLC. This software builds Debruyne graph like data structures on which it aligns the longer (PACBio or Illumina reads). In our case, given we had MiSeq reads of 300nt, we used DBG2OLC to extend the existing contigs from the CLCBio step. This result was then scaffolded further using DBG2OLC again, with the PACBio reads. Given the parameters are different depending the input data, both could not be done simultaneously. To finish the assembly we ran once more DBG2OLC, again with the MiSeq reads to elongate the obtained large contigs from the previous step.

8.4.5 Genotyping-by-Sequencing

Leaf samples of the parents of the 366 lines were collected and stored at -80°C. Prior to extraction, leaf samples were grinded using a Retsch Mixel Mill MM 400. gDNA were extracted using NucleoSpin Plant II mini column (Macherey Nagel) according to manufacturer's instructions starting from 100 mg of leaf tissues powder.

The Genotyping-by-Sequencing (GBS, Elshire et al. 2011) procedure was subcontracted to Biogenomics, a KULeuven (Leuven, Belgium) R&D subdivision specialized in project management and consulting, product development and services in animal and plant genomics. The GBS procedure is described in the next paragraph.

Library preps were achieved according to the Elshire et al. (2011) protocol using the ApeKI restriction enzyme. These libraries were Pair-end sequenced in 5 lanes of an Illumina Hiseq 2000 sequencer. GBS sequencing reads were demultiplexed using a custom-made python script. Pair-end reads were merged using PEAR (Zhang et al. 2014). The reads were mapped on the available reference genome using default settings of BWA (Li and Durbin, 2009). Non mapping reads and reads mapping on more than one location were removed. SNP and indels calling was done using Freebayes (Garrison and marth, 2012). Only variants present in at least 10% of individuals and with a minor allele frequency above 0.05 were retained. A sample was considered as homozygous if at least five reads supported it. An heterozygous sample was always considered valid, even if it had only one read for both the reference and the alternate allele. Raw SNPs were further filtered to retain SNPs with less than 30% missing data and a minor allele frequency greater than 0.05 using a custom R script. Finally, the missing genotypes were imputed using random forest regression imputation method implemented in the R package “MissForest” (Stekhoven and Bühlmann, 2011).

8.4.6 Population analysis

The population structure was analyzed by principal component analysis (PCA), assignment method and analysis of molecular variance (AMOVA). These methods were applied both on the 15 SSRs data set and on a subset of

1000 high quality SNPs (less than 1 % missing data, a minor allele frequency higher than 0.2 and one SNP per contig of the reference genome).

PCA were performed with R 2.14.0 software (R Development Core Team 2013), using functions in the packages *ade4* (Jombart, 2008) and *ade4* (Chessel et al. 2004). The assignment method was applied with the software *Structure* (Prichard et al. 2000) using the admixture model. Different numbers of putative populations (K) were tested, ranging from 2 to 12. Ten independent runs were carried out per K and 10^6 Markov chain Monte Carlo iterations were used after a burn-in of 10^5 steps. The best number of putative population was evaluated with the Evanno method and was 8 (Evanno et al. 2005). AMOVA analyses and F_{st} values were computed with Arlequin version 3.11 software (Excoffier and Lischer 2010)

The genetic diversity was evaluated as the number of alleles, and Nei's diversity index (I_{Nei}), computed as $I_{Nei} = 1 - (\sum \sum p_{ij}^2) / n$, where p_{ij} is the frequency of the i th allele of the j th locus (with a total number of n loci). Those parameters were computed with custom R scripts.

The matrix of familial relationships (K) used in the linkage disequilibrium (LD) and GWAS analyses was estimated with the same 1000 SNPs according to Van Raden (2008).

8.4.7 Linkage disequilibrium

The degree of linkage disequilibrium (LD) between all possible pair of SNPs present on the same contig was quantified with the parameter r^2 (Hill and Robertson 1968) obtained with and without taking into account the population structure evaluated by assignment method (Q) and familial relatedness (K) following Mangin et al. (2012). Estimates of r^2 , accounting for population structure and familial relatedness, were obtained using the R package LDcorSV (Mangin et al. 2012).

8.4.8 GWAS

GWAS was performed using a compressed mixed linear model (cMLM) implemented in the R package GAPIT (Zang et al. 2010, Liptka 2012). Four models were tested. The first model included corrections accounting for the familial relatedness and structure results (structure model). The second model included correction for the familial relatedness and the PCA results (PCA model). The ideal number of principal components was evaluated according to the BIC criterion. The third model was a naive model with no correction for population and/or familial relatedness (Naive model). And the last model (FEH model) was a model accounting for familial relatedness, population structure evaluated by PCA and accounting for the effect of the FEH indel described in Dauchot et al. 2015. Quantile-quantile plots were produced to assess the extent to which the analysis produced more significant results than expected by chance.

We used $\alpha = 5.1 \times 10^{-6}$ as the genome-wide significance threshold for the full population, which corresponds to $\alpha = 0.05$ after Bonferroni correction for 9859 SNPs.

8.5 Results

8.5.1 Reference genome

The sequencing of the reference individual by the illumina-sequencing platforms generated approximatively 141 Gbases of raw data, corresponding to a 100× coverage. These sequences were assembled into 21 680 non-redundant contigs covering a total of 1.1 Gb of the 1.3 Gb chicory genome. The N50 was 6505 bp. This assembly is a preliminary step of a genome sequencing project (Dauchot et al. unpublished data).

8.5.2 SNPs discovery and distribution

GBS generated 500 10⁹ million reads for the 366 individuals. Finally, 122 650 SNPs could be mapped on 11 312 contigs of the reference genome. A high fraction of these SNPs presented a high proportion of missing data (Supplementary material 17 and Supplementary material 17). A subset of 9 859 SNPs mapped on 2 690 contigs, presenting a calling rate and a minor allele frequency superior to 0.66 and 0.05, respectively, was selected for the genome-wide association mapping study (GWAS) and the linkage disequilibrium (LD) analysis. 1741 contigs included more than one SNP of the subset (Figure 46). Most of the successive SNPs from the subset (90%) were separated by less than 400 bp from each other (Figure 46 and Figure 46), which corresponds to the size of the DNA fragment of GBS libraries

(170-350 bp, Elshire 2011). SNPs appeared to be randomly distributed on the reference genome.

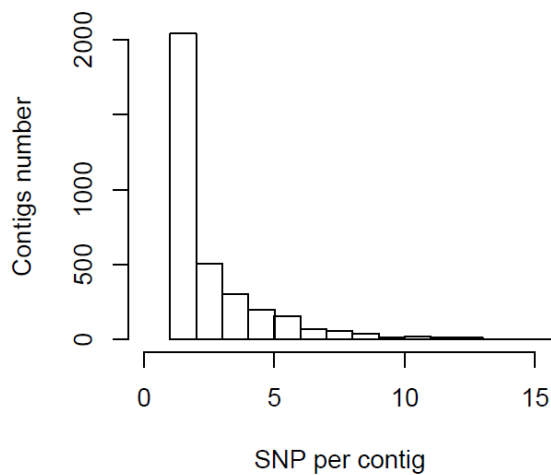


Figure 45 : Histogram of the number of SNP per contig

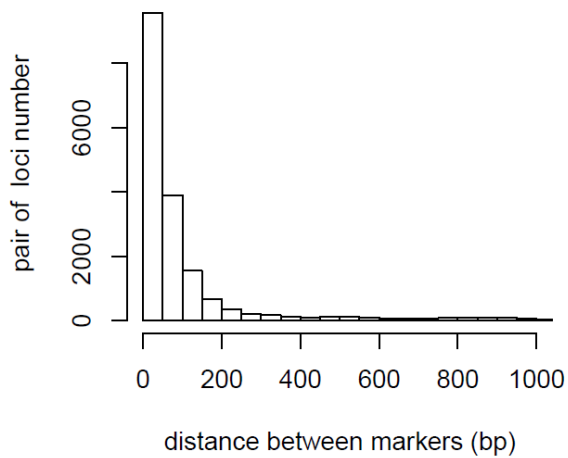


Figure 46: Histogram of distance between pairs of SNPs

8.5.3 Population structure and linkage disequilibrium

The diversity panel of 366 individuals was sampled out of a collection of 1384 individuals from 15 cultivars previously described in Raulier et al. 2015. The panel presented 230 out of the 263 SSR alleles present in the collection and the same genetic diversity index ($I_{Nei} = 0.9$). The PCA, Structure results, ANOVA and Nei genetic index ($I_{Nei} = 0.9$) of the collection and the panel were very similar, indicating that the population structure of the original collection was conserved in the panel. The PCA (Figure 47) and Structure results (data not shown) of the panel obtained from the SSR markers and the 1000 SNPs were also very similar. However, the cultivar effect explained less genetic variance with the AMOVA based on 1000 SNPs (6%) than with the AMOVA based on 15 SSR markers (16%). The population consisted of one homogenous group including the nine French cultivars, of a smaller homogenous group including the Belgian cultivars ‘Dageraad’ and ‘Halle’ and of cultivars clearly differentiated: ‘Novipa’ (Belgium), ‘Polanowicka’ (Poland), ‘Kujawska’ (Poland) and ‘Fredonia’ (Austria). ‘Novipa’ was the most differentiated cultivar, while ‘Fredonia’ showed some similarities with the French group and other cultivars.

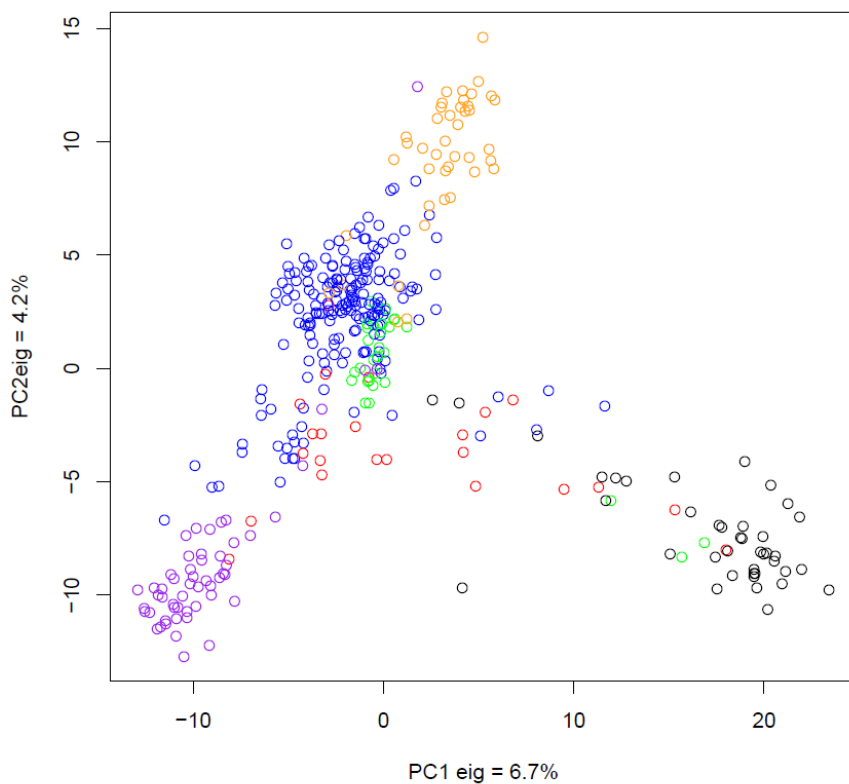


Figure 47 : PCA of the 366 individuals of the panel based on 1000 SNPs. Blue: Pévèle, Fruitosa, Cassel, Orchies, Tilda, Sabau 3, Bergues, Candi, and Hicor, red: Fredonia, green: Polanowicka, orange: Kujawska, purple: Dageraad and Halles, black: Novipa.

The linkage disequilibrium (LD) fell below 0.2 at 350 bp and remained between 0.1 and 0.2 between 350 and 3000 bp (Figure 48). The average LD estimated on 50 bp intervals followed a non-linear trend.

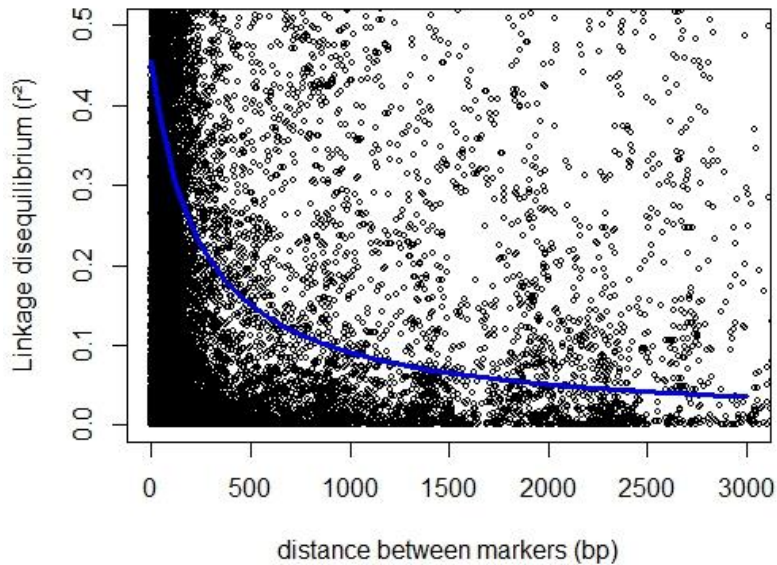


Figure 48: Linkage disequilibrium (r^2) among marker pairs as a function of physical distance.

8.5.4 Phenotyping

A total of 112, 122 and 139 chicory lines were phenotyped during the trials 1, 2 and 3. At harvest, the average dry weigh (DW), dry matter content (DMC) and carbohydrate content (CC) were 228g, 24.6 % and 17.8 %, respectively

(Figure 49). The carbohydrates mainly consisted of inulin (IN = 95.2 %) and, in a lower proportion, of sucrose (SUCn = 4.3 %). The free fructose (FFn) and free glucose (FGn) represented 0.29 and 0.17 % of the carbohydrates content, respectively. The average degree of polymerization (DPin) was 12.3. After the exposure to low temperatures, the following significant variations were observed: (i) a small decrease of DMC (-0.5 %), (ii) a small increase of CC (+1.6%) (iii) a strong reduction of DPin (-5) and IN (-18.2%), (iv) a high increase of SUCn (+13%) and FFN (+4.6%) and (v) a slight increase of FGn (+ 0.27).

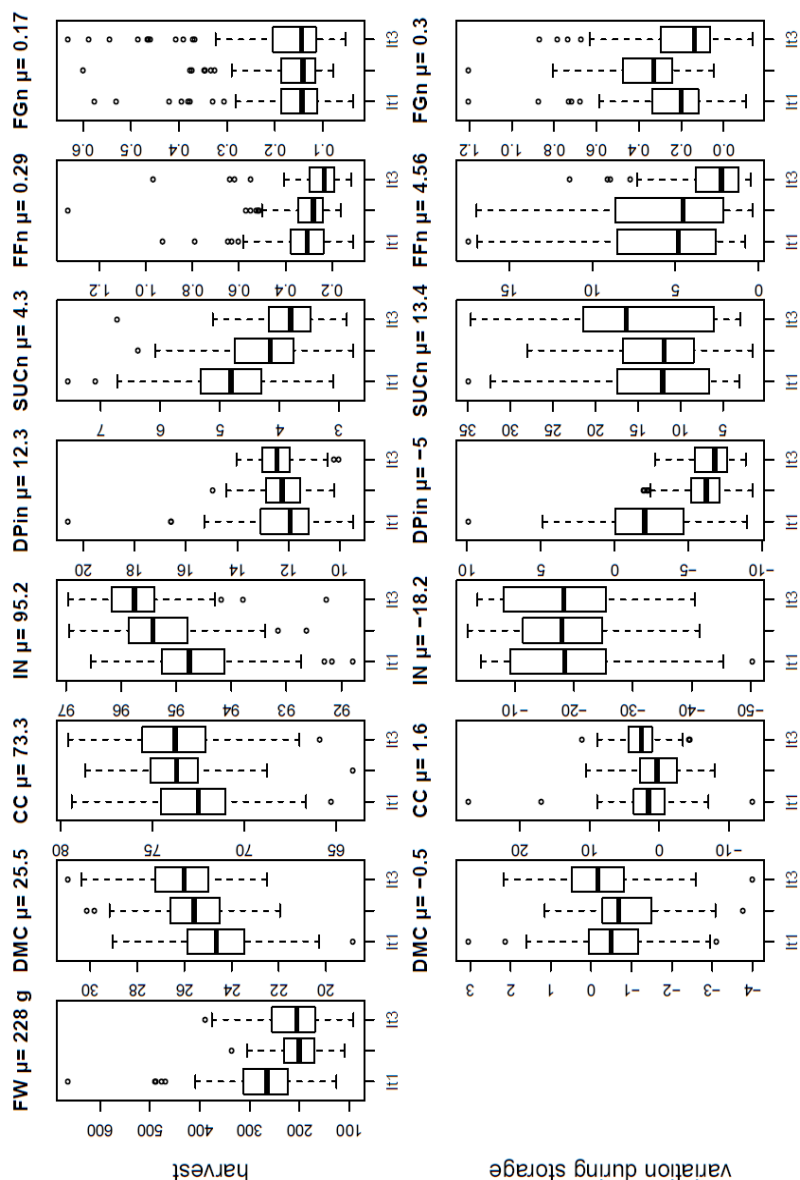


Figure 49: boxplots presenting the result of the trials 1, 2, 3 (from left to right) of the fresh weigh (FW), dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), free glucose content (FGn), free fructose content (FFn) and sucrose content (SUCn), at harvest and their variation after cold exposure.

The trials presented some significant differences for 6 variables (Table 16 and Table 17) and the cultivars presented some significant differences for most variables. In particular, the trial effect explained up to 20% of the variability of fresh weight, IN and SUCn at harvest and up to 43% of the variability of variation of DPIn after cold exposure. The proportions of the phenotypic variability explained by the cultivar (Table 16 and Table 17) and the heritabilities (Table 18) were in the same range (0 to 0.4). The effect of the interaction trial×cultivar was significant only for the variation of DPIn after cold exposure. Therefore, the variables presenting a significant trial effect were normalized by retrieving the trial effect in such a way that, for each cultivar, the mean in the different trials was the same. These normalized data were used to compute all results presented below.

Harvest								
	FW	DMC	CC	IN	DPin	FGn	FFn	SUCn
p-val trial	0,0001	0,0017	0,0453	<.0001	0,50	0,4446	0,1201	<.0001
p-val cult	0,0002	<.0001	<.0001	0.0002	0,04	0.0187	0.3111	<.0001
					5			
p-val trial*cult	0,21	0,72	0,78	0,93	0,26	0,90	0,76	0,81
R² trial*cult	0,41	0,41	0,28	0,38	0,12	0,20	0,23	0,40
R² var	0,18	0,29	0,20	0,17	0,05	0,16	0,17	0,16
R² trial	0,21	0,11	0,04	0,24	0,01	0,01	0,07	0,26

Table 16:P-values (p-val) and R² of the trial, cultivar (cult) and the trial-cultivar interaction (trial*cult) effects on of the following variables measured at harvest: fresh weight (FW), dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), Free glucose content (FGn), Free fructose content (FFn) and sucrose content (SUCn).

Variation after cold exposure							
	DMC	CC	IN	DPin	FGn	FFn	SUCn
p-val trial	0,1164	0,1742	0,481	<.0001	0,0909	0,0067	0,1648
p-val cult	0,0487	0,7319	<.0001	<.0001	0,0102	<.0001	<.0001
p-val trial*cult	0,22	1,00	0,84	0.0051	0,77	0,63	0,51
R² trial*cult							
R² var	0,19	0,16	0,36	0,60	0,27	0,37	0,42
R² trial	0,09	0,10	0,32	0,20	0,20	0,28	0,37
p-val trial	0,08	0,10	0,00	0,43	0,13	0,15	0,02

Table 17: P-values (p-val) and R² of the trial, cultivar (cult) and the trial-cultivar interaction (trial*cult) effect on the variation after cold exposure of the following variables: dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), Free glucose content (FGn), Free fructose content (FFn) and sucrose content (SUCn).

h²	FW	DMC	CC	IN	DP	FGn	FFn	SUCn
Harvest	0.002	0.152	0.243	0.126	0.054	0.274	0.252	0.089
Variation after cold exposure		0.042	0.232	0.396	0.300	0.196	0.323	0.388

Table 18: Heritability (h²) of the following traits, measured at harvest and the variation of these traits after cold exposure : fresh weight (FW) dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), Free glucose content (FGn), Free fructose content (FFn) and sucrose content (SUCn) measured at harvest and their variation after cold exposure.

8.5.5 GWAS

The GWAS was initially performed with 9859 SNPs on all described variables with three different mixed models. The first model included a correction for the familial relatedness (K) and population structure (Q) evaluated by Structure (Structure model), the second model included a correction for the familial relatedness (K) and population structure (Q) evaluated by PCA (PCA model) and the third model included no correction (Naive model). QQ plots produced by the PCA model were slightly better than that of the Structure model. Both of them yielded a much better fit to the theoretical distribution than the Naive model (Supplementary material 21 and Supplementary material 22). The PCA model was used for the subsequent analyses.

Most variables measured at harvest were highly correlated to each other and the variations of the carbohydrates-related variables (IN, DPin, FGn, FFn, SUCn) after cold exposure were highly correlated to each other (Supplementary material 18, Supplementary material 19, Supplementary material 20).

The significant results of the PCA model are given in the Table 19. Respectively one, two, two and one different SNPs were significantly associated to FW, DMC, DP and FFn. These SNPs accounted for 3 to 4 % of the phenotypic variability of these variables. At harvest, the population structure and family relatedness explained between 20 and 40 % of the phenotypic variability of yield-related variables (FW, DMC and CC) and between 3 and 7 % of the phenotypic variability of carbohydrates-related variables (IN, DPin, FGn, FFn and SUCn) (Table 20). Eleven SNPs and the indel in FEH gene described in Dauchot et al. 2014 were significantly or highly associated with the variation of carbohydrates-related variables after cold exposure (IN, DPin, FGn, FFn, and SUCn). The population structure and family relatedness explained 3 % of the phenotypic variability of the variation of DMC and CC after cold exposure and between 12 and 27 % of phenotypic variability of the variation after cold exposure of carbohydrates-related variables (IN, DPin, FGn, FFn, and SUCn).

Except for one SNP (53713_4525), all SNPs associated to the variation of a trait after cold exposure were in high LD (mean $r^2 = 0.48$, Figure 50) and their association did not remained significant when the effect of the FEH indel was taken into account (data not shown). These SNPs were dispatched

on different contigs of the reference genome, covering together 360 Kb. Therefore, we expect that (i) these SNPs and contigs are physically close to the genomic region including the three 1-FEH genes and (ii) the LD in this region extends on long distance compared to the LD estimated on the whole genome. The SNP 53713_4525 was in low LD with the other significant SNPs (mean $r^2 = 0.06$, Figure 50) and its association with the variation after cold exposure of sucrose content remained significant when taking into account the FEH indel, explaining up to 8% of the phenotypic variance of inulin content. Therefore we may expect that this SNP corresponds to a new QTL associated to the susceptibility to inulin degradation after cold exposure.

GWAS analysis of carbohydrate and inulin quality in chicory

		Harvest								Variation after cold exposure							
Contig	SNP	FW	DM	CC	IN	Dpin	FGn	FFn	SUCn	DM	CC	IN	Dpin	FGn	FFn	SUCn	
15097	271	5.8	0.7	18	2.0	0.2	0.5	0.6	2.1	0.5	0.1	0.1	0.6	0.2	0.2	0.0	
15097	287	0.3	0.1	0.2	0.0	0.2	0.1	0.4	0.0	0.2	0.3	0.1	0.0	0.1	0.2	0.0	
15097	264	0.3	0.5	0.3	0.4	0.8	0.5	0.6	0.4	0.0	0.6	0.3	0.7	0.4	0.2	0.2	
50696	11385	0.0	0.8	0.7	0.4	0.0	0.1	0.3	0.7	0.3	0.1	1.0	0.6	1.6	0.9	0.7	
50696	9743	0.3	0.9	0.9	0.1	0.1	0.7	1.2	0.4	0.0	0.3	0.6	0.1	0.9	0.6	0.3	
50696	9211	0.6	5.3	0.5	0.5	0.5	0.8	0.9	0.6	1.6	0.4	0.9	0.0	0.9	1.3	0.5	
110154	7693	0.4	3.6	1.5	1.6	1.7	0.8	0.2	1.5	0.5	0.1	2.9	0.2	1.4	1.7	2.5	
110154	7686	0.3	3.1	0.8	1.8	2.1	1.1	0.2	1.8	0.1	0.3	2.0	0.1	0.4	1.1	1.8	
110154	7659	0.0	2.6	0.5	1.6	1.4	0.5	0.1	1.4	0.6	0.2	2.4	0.2	0.7	0.6	2.3	
110154	7642	0.7	0.3	0.2	0.7	0.1	0.3	0.9	0.4	0.1	0.4	0.1	0.8	0.1	0.4	0.0	
110154	7639	1.8	5.2	0.3	1.1	0.8	0.6	0.5	0.9	0.0	0.3	1.9	0.2	1.0	0.7	2.0	
279	99	0.2	0.6	0.0	0.3	0.6	0.2	0.1	0.2	0.8	0.2	0.2	0.3	0.4	0.1	0.5	
279	110	0.4	1.8	0.1	1.5	6.0	1.1	0.8	1.1	0.0	0.3	1.2	1.5	1.2	2.4	0.4	
279	123	0.4	0.3	0.1	0.2	0.5	0.2	0.1	0.1	0.1	0.6	0.3	0.2	0.5	0.2	0.4	
902	2115	0.3	1.1	1.0	0.7	0.0	0.0	0.4	0.6	0.2	0.4	0.1	0.5	0.1	0.5	0.1	
902	2178	1.2	0.8	0.7	1.1	0.1	0.9	5.1	0.9	0.1	0.3	0.1	0.0	0.1	0.0	0.1	
902	2196	0.0	0.8	0.6	0.6	0.1	0.0	0.3	0.5	0.2	0.6	0.1	0.5	0.1	0.4	0.0	
902	2205	0.0	0.1	0.1	0.1	0.5	0.0	0.0	0.2	0.0	1.7	0.2	0.5	0.7	0.5	0.1	
90371	4373	0.2	1.2	0.4	0.1	0.3	0.5	0.3	0.1	1.0	0.8	6.2	3.0	3.4	2.6	6.9	
90371	4424	0.3	0.5	0.1	0.2	0.1	1.0	0.3	0.0	0.3	0.0	0.2	0.0	0.1	0.0	0.3	
90371	4442	0.4	0.7	0.1	0.1	0.1	0.7	0.1	0.1	0.2	0.0	0.5	0.0	0.1	0.0	0.5	
90371	4464	0.2	0.3	0.1	0.1	0.0	0.9	0.2	0.1	0.7	0.0	0.3	0.0	0.1	0.0	0.4	
90371	4472	0.3	0.6	0.1	0.2	0.2	0.9	0.3	0.0	0.5	0.0	0.6	0.1	0.0	0.1	0.6	
90371	4487	0.2	0.5	0.2	0.1	0.1	0.7	0.7	0.3	1.3	0.8	5.0	4.0	4.1	3.7	4.6	
FEH indel		0.1	0.3	0.4	0.0	0.2	1.1	1.5	0.3	5.6	2.9	27.6	16.6	12.7	14.5	27.2	
331898	190	0.3	0.3	0.5	0.7	0.2	0.5	0.1	1.0	2.8	0.9	8.2	5.3	4.4	2.9	9.3	
122289	342	1.8	0.1	0.2	1.0	0.2	0.4	1.0	1.2	0.8	0.7	2.9	1.9	1.5	3.5	2.2	
122289	305	1.1	0.1	0.0	0.8	0.5	0.1	0.8	1.1	0.6	0.8	1.9	1.4	0.6	2.3	1.4	
122289	1947	1.4	0.3	0.2	1.1	0.4	0.2	0.5	1.7	2.2	1.2	8.2	5.9	4.3	5.9	7.4	
122289	1843	2.0	0.5	0.2	0.5	0.9	0.3	0.2	0.7	0.7	0.6	5.7	2.0	2.9	4.6	4.7	
122289	2174	1.0	0.2	0.7	1.0	1.2	0.1	0.3	1.4	2.0	1.0	9.4	3.8	3.6	6.3	8.7	
122497	2323	0.6	0.1	0.1	0.5	0.7	0.4	0.1	0.4	0.9	0.3	1.9	2.6	1.5	0.3	2.3	
122497	2374	1.2	0.0	0.8	0.5	0.5	0.2	0.0	0.8	1.8	0.7	8.5	4.1	4.7	5.9	7.6	
79224	504	0.2	0.2	0.5	0.0	0.3	1.4	0.3	0.4	1.8	0.9	8.1	3.2	6.1	5.0	7.6	
79224	528	0.8	0.1	0.5	0.2	0.5	2.2	1.2	0.1	1.9	1.1	11.0	4.7	5.7	7.6	9.4	
79224	534	0.3	0.5	0.6	0.3	0.0	2.8	0.9	0.0	1.3	0.5	7.4	3.1	4.1	4.6	6.8	
68436	1089	1.1	0.2	0.7	0.2	0.5	0.3	0.2	0.5	1.7	1.4	9.2	5.5	4.4	6.2	8.4	
53713	4525	0.2	1.5	0.6	0.0	0.0	0.3	0.1	0.1	1.3	0.0	9.2	3.8	4.1	3.8	9.1	

Table 19: $-\log_{10}$ of the p-value for the SNPs presents on contig including at least one SNP presenting at least a significant association for at least one of the following parameters: fresh weight (FW), dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), Free glucose content (FGn), Free fructose content (FFn) and sucrose content (SUCn) measured at harvest and their variation after cold exposure. The length of the color bar is proportional to the $-\log_{10}$ of the p-value. The red bars correspond to variables measured at harvest and the blue to the variation of these variables after cold exposure. The first and second column includes the contig name and the second column the position of the SNP on the contig.

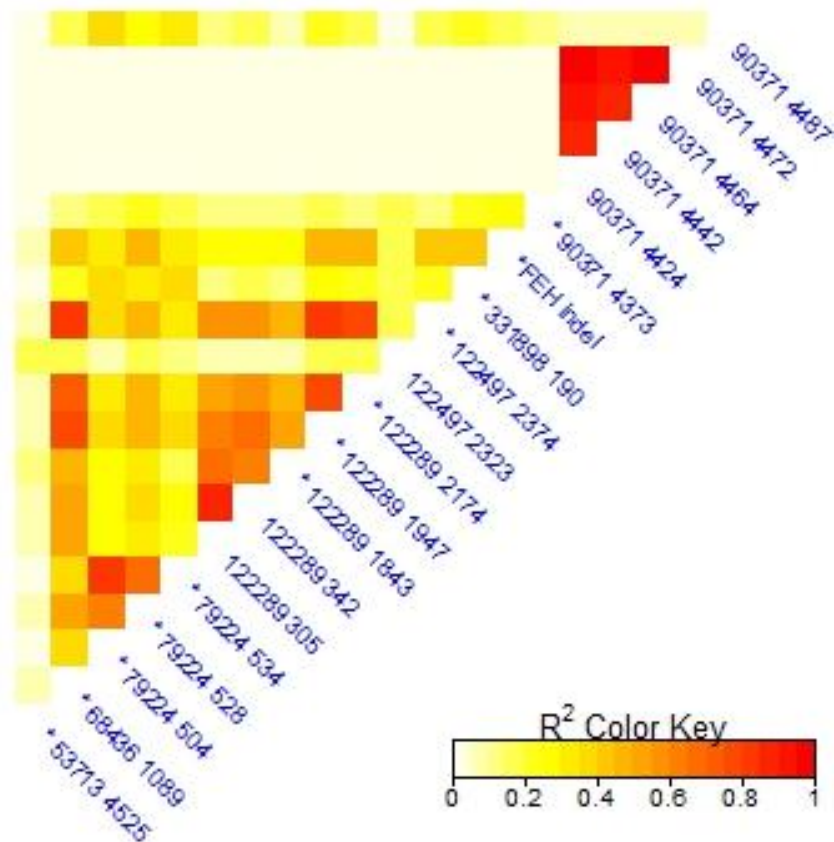


Figure 50: LD heat map of the SNPs present on the contigs including SNPs with significant association to the variation of a trait after cold exposure. The labels of SNPs consist in the contig number followed by the SNP position on the contig. The SNPs presenting significant association with a trait are marked with a star (*).

		FW			DMC			CC			IN			DP			FGn			FFn			SUCn					
		Number of SNPs			R ² _{SNP} (%)			R ² _{pop + kinship} (%)			Number of SNPs			R ² _{SNP} (%)			R ² _{pop + kinship} (%)			Number of SNPs			R ² _{SNP} (%)			R ² _{pop + kinship} (%)		
Harvest		1	6	10	2	3	39	0	/	20	0		9	2	4	3	0	/	12	1	3	13	0	/	7			
Variation after cold exposure					1	3	3			3	11	25	24	4	16	15	3	13	6	4	15	12	11	34	27			

Table 20: Number of SNPs presenting a significant association, proportion of phenotypic variability associated to the SNPs with the lowest p-value (R²_{SNP}), and the proportion of phenotypic variability explained by the population structure (Q) and familial relatedness (R²_{Q+K}), for the following variables : fresh weight (FW), dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), free glucose content (FGn), free fructose content (FFn) and sucrose content (SUCn) measured at harvest (H) and their variation after cold exposure .

8.6 Discussion

In chicory, inulin yield and quality at harvest and the sensitivity to inulin degradation induced by cold temperatures determine the quality of the harvest. Previous studies identified an indel in an active site in 1-FEH IIb associated to 35 % of the phenotypic variability of the sensitivity of inulin depolymerisation after exposure to low temperature (Dauchot et al. 2014, 2015). Investigations of the polymorphisms detected in 9 other genes of the family 32 of glycoside hydrolases (GH32), including 1-FFT and 1-SST, did not show significant association between these polymorphisms and variables describing inulin quality at harvest or following degradation. The present study investigated a diversity panel representative of the cultivars at the origin of the current cultivars of industrial chicory. Using this panel, characterized by a very low extent of the linkage disequilibrium (LD), we identified several QTLs associated to carbohydrates and inulin characteristics and the sensitivity to inulin degradation and one major QTL corresponding to the 1-FEH IIb.

The panel of individuals required in genome-wide association studies (GWAS) may either consist in elite material or a diversity panel (Breseghello et al. 2006). The case of chicory is particular because the breeding of industrial chicory for inulin production only started in the '80s with the reconversion of sugar beet refinery into chicory refinery in Belgium and the Netherlands. Therefore, the germplasm at the origin of the current cultivars is very close to the elite material and combines the following characteristics: (i) it includes all the diversity present in elite material, (ii) it is phenotypically close enough to the elite material to

easily introgress valuable characters (iii) its LD extent is expected to be lower. We sampled our diversity panel in this material with a method originally implemented to select core collections. This method aims at selecting individuals that represent the original collection by decreasing the redundancy among individuals and is supposed to decrease the population structure and the LD among unlinked loci (Marita et al., 2000, Jamann et al., 2015). In our study, the sampling method was successful to sample a panel of 366 individuals that equally represented the collection of 1297 individuals described in Raulier et al. (2015) in terms of population structure and genetic diversity. The population structure and familial relatedness had little influence on the GWAS analysis as the SNPs with lowest p-values in the different models were the same (data not shown). The main differences between models were a higher deviation in QQplot and the lower p-value in the Naive model. Therefore, the Naive model is expected to lead to a higher false discovery rate.

The sequencing of the chicory genome was initiated to complete the genomic resources available in chicory and to help the identification and mapping of the SNPs generated by GBS. The successful Genotyping-by-Sequencing of chicory revealed that this technique is very appropriate to non-model species with low genomic resources. Among the 122 000 SNPs identified, 9859 presented a sufficient minor allele frequency and calling rate to be used in the GWAS. The remaining SNPs may be useful to increase the number of markers in regions of interest in our panel or in other populations.

The LD defines the association between loci and, in the context of a GWAS delineates the size of DNA fragments explored by a single genetic marker.

According to Ersoz et al. (2007), the minimal distance to detect an association between a marker and a QTL corresponds to the distance at which average r^2 is reduced to 0.10. The reasoning for this r^2 cut-off is as follows: in a complex trait, a large QTL may explain approximately 10% of the phenotypic variance, a marker presenting a LD of $r^2=0.1$ with this QTL, only explains 10% of the total QTL variation, namely one percent of the phenotypic variance. Some authors increase this cut off up to $r^2=0.8$ (Barrett 2006, Bastien et al. 2014). The chicory panel was characterized by a low LD extent $r^2 < 0.2$ at 300 bp and $r^2 < 0.1$ at 900 bp. Such a low LD extent was described in other open pollinated species such in rye, maize, ryegrass and douglas pine (Li et al. 2011, Tenaillon et al. 2001, Xing et al. 2007, Krutovsky and Neale, 2005). It implies that the GWAS on these populations could result in the identification of very accurate localization provided a very high SNP density. Taking a r^2 threshold of 0.1 or 0.2, GWAS of our panel should require 144 000 000 or 433 000 000 markers respectively. These values almost correspond to the resequencing of the genome, which is not currently reasonably achievable for a non-model species. Another genotyping alternative is to genotype a very large panel of candidate genes such as realized in *Lolium perenne* (Rutting et al 2015). The identification of such SNPs could be generated by targeting sequencing of candidate genes. The reference genome could be a useful tool in this context. Another alternative would be to implement GWAS on population presenting a larger LD extent such as the elite lines or experimental populations.

According to the method described above, with a r^2 threshold of 0.1, our SNP set explored 0.07 Gb of the 1.3 Gb genome, which is a small genome fraction (Koopman, 2002). Nevertheless, this SNP set captured at least 6 QTLs and

among them, one major QTL corresponding to the 1-FEH IIb loci. Different factors contributed to the high QTL yield compared to the theoretical need of markers. Firstly ApeKI, the restriction enzyme used in our GBS protocol, is methyle sensitive so that it cleaves at restriction sites with no methylated cytosine residues. Since most highly repetitive regions of the genome are highly methylated, ApeKI has the advantage of enriching for lower-copy regions which tend to carry most of the genes associated with functional variation (Elshire et al. 2011, Jamman et al. 2015). Therefore the explored regions of the genome tend to be the most informative in term of genetic determinism. Secondly, the LD is unevenly distributed on the genome (Flint-Garcia, 2003) and extends on larger distance in genomic sections including loci under selection (Flint-Garcia, 2003, Prasad et al. 2008). In the present study, LD surrounding the major QTL corresponding to the 1-FEH IIb gene extends on long distance compared to the LD estimated on the whole genome. It is ranging between 0.15 and 0.89 on a window of 100 Kb. The particularly high extent of LD surrounding the 1-FEH IIb gene may reflect a selection event against inulin breakdown combined to hitchhiking mechanism (Ohta and Kimura 1975). Therefore, one may expect that LD around QTLs that undergone selection pressure (e.g. major QTLs) extends on a larger distance than the genome-wide LD. The identification of such QTL may be facilitated because it requires a lower marker density compared to the estimation based on genome-wide LD.

The plants of the different trials were harvested between the end of September and the beginning of October. Therefore, they were in the second phase of inulin metabolism described by Van Arkel et al. (2012). At this time, the carbohydrate and fructan profiles result from both (i) the combined synthesis activities 1-SST

and 1-FFT occurring during the first phase and in the beginning of the second phase and (ii) the 1-FFT back-transfer activity occurring during the second development phase. The highly negative correlation between inulin content (IN) and sucrose content (SUCn) may reflect the fact that, at the beginning of the second development phase, (i) the sucrose is not a limiting factor and (ii) the variability in inulin content depends on the ability of the plant to consume available sucrose to generate inulin, which is the function of the 1-SST during the first and second phases and of 1-FFT during the second phase.

At harvest, one, two, two and one SNPs with minor effect (R^2 of 3 to 4), all located on different contigs were significantly associated to the fresh weight (FW), dry matter content (DMC), polymerization degree (DP) and fructose content (FFn), respectively. A better assembly of the reference genome will determine if the SNPs identified for DMC or DP corresponds to one or two QTLs. Moreover, an investigation of the genomic regions surrounding the significant SNPs with higher SNPs density could increase the accuracy and the power of the QTL identification and determine the actual effect of the QTL.

One SNP significantly associated to DP showed some relatively low (although non-significant) p-value for sucrose content (SUCn) and, to a lesser extent, for inulin concentration (IN). This SNP may be physically close to a QTL affecting DP, sucrose and inulin content characters.

A part of the harvested roots were exposed to low temperatures to artificially trigger the 1-FEH activity that characterizes the third phase of development described by Van Arkel et al. (2012). After the cold exposure, inulin degradation

was reflected by the decrease in mDP and inulin content for the benefit of an increase in glucose, fructose and sucrose content. These parameters were highly correlated to each other and presented high association with 11 SNPs located on 6 different contigs and to the indel located in FEH described in Dauchot et al. (2014). Most of these SNPs correspond to the 1-FEH IIB QTL while the SNP 4525 present on the contig 53713 may correspond to a new QTL for sensitivity to inulin degradation. Considering the low heritabilities ($0.196 < h^2 < 0.396$) and the occurrence of a major QTL for the variables describing variation of carbohydrates after cold exposure, no other major QTL is expected for inulin degradation.

The fructan profile in segregating populations of wheat, onion and ryegrass has been shown to follow one or few major QTLs and possibly some QTLs with minor effect. The major QTLs seem to influence the regulation of expression of fructan genes such as the *1-SST*, *1-FFT* and *6-SFT* cluster and *6-FEH* in wheat and possibly *1-SST* in onion (Huynh et al. 2012; Zhang et al. 2008a, Müller et al. 2000; Wang et al. 2000, Müller et al. 2000). The chicory case is different since the 1-FEH IIA indel, which affects fructan sensitivity to inulin degradation after cold exposure, corresponds to the loss of function of an enzyme. Among the minor QTLs detected in others species, only one candidate gene has been identified (*LpCAD2*, in rye grass) and is associated with stem elongation and therefore may influence the carbohydrate source-sink relationship. In chicory, the QTLs identified for fructan profile at harvest could either (i) involve pleiotropic effects associated to sink-source mechanisms, or (ii) influence the expression of a fructan gene alike the major QTL identified in onion (Müller et al. 2000; Wang, Van den Ende, and Tillberg 2000), or (iii) affect the enzyme affinity for

acceptor of fructosyl unit. Indeed, the affinity of 1-FFT for low DP or high DP fructosyl acceptor and the 1-FEH affinity to low DP or high DP inulin are at the origin of different inulin profiles among different species (Itaya et al. 2007, Tamura et al. 2009, Vergauwen et al. 2003). The sink-source pleiotropic effect potentially includes many different process, ranging from the efficiency of the photosynthetic apparatus, the strategy of the plant to transfer assimilates to different organs, followed by the repartition of these assimilate among the different organs. Moreover, many parameters, from the molecular level to the general architecture of the plant may influence these processes. Therefore, many physiological processes may potentially influence the flux of assimilate in the root vacuole, where the fructan synthesis occurs. These potentially complex mechanisms, combined to the low heritabilities and absence of identified major effect QTL, point toward a complex multigenic determinism of inulin and carbohydrate content.

The absence of major effect QTL for inulin yield quality at harvest in chicory probably reflects a selection toward yield and inulin quality at harvest in the tested panel, that fixed the favorable alleles. Conversely, the occurrence of major effect QTL for inulin degradation in industrial chicory, fructan content in wheat, onion and ray grass (Huynh et al. 2008, 2012; Zhang et al. 2008a McCallum et al. 2006; Raines, Henson, and Havey 2009; Yaguchi et al. 2008, Turner et al. 2006) may reflect the absence or low selection against these traits in the studied material.

In conclusion, the investigation of our panel highlighted the complexity of the genetic determinism of carbohydrates in chicory. GWAS yielded positive results

considering the low extent of LD. GWAS in low LD extent population results in the requirement of high number of markers, but the advances in genotyping technologies should overcome this limitation within a few years. Moreover, we observed that major QTLs present a larger LD extent compared to the genome-wide LD, so that they may be detected with a lower marker density.

Conflict of interest

The authors acknowledge the Region wallone (DGARNE-Belgium) for supporting this research (grant D31-1221).

Conflict of interest

O. Maudoux and C. Notté are employees of Cosucra-Groupe Warcoing S.A.

8.7 References

- Amaducci, S., Pritoni, G. (1998) Effect of Harvest Date and Cultivar on Cichorium Intybus Yield Components in North Italy. *Industrial Crops and Products* 7: 345–49.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4.[Computer software manual]. R package version 1.1-7.
- Baert, J. R. (1997). The effect of sowing and harvest date and cultivar on inulin yield and composition of chicory (*Cichorium intybus* L.) roots. *Industrial Crops and Products*, 6(3), 195-199.
- Barrett, J. C., & Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature genetics*, 38(6), 659-662.
- Bastien, M., Sonah, H., & Belzile, F. (2014). Genome wide association mapping of resistance in soybean with a genotyping-by-sequencing approach. *The Plant Genome*, 7(1).
- Blecker, C., Fougny, C., Van Herck, J. C., Chevalier, J. P., & Paquot, M. (2002). Kinetic study of the acid hydrolysis of various oligofructose samples. *Journal of Agricultural and Food Chemistry*, 50(6), 1602-1607.
- Breseghello, F., & Sorrells, M. E. (2006). Association analysis as a strategy for improvement of quantitative traits in plants. *Crop Science*, 46(3), 1323-1330.
- Chessel, D., Dufour, A. B., & Thioulouse, J. (2004). The ade4 package-I-One-table methods. *R news*, 4(1), 5-10.
- Dauchot, N., Raulier, P., Maudoux, O., Notté, C., Bertin, P., Draye, X., & Van Cutsem, P. (2014). Mutations in chicory FEH genes are statistically associated

with enhanced resistance to post-harvest inulin depolymerization. *Theoretical and applied genetics*, 127(1), 125-135.

Dauchot, N., Raulier, P., Maudoux, O., Notté, C., Draye, X., & Van Cutsem, P. (2015). Loss of function of 1-FEH IIb has more impact on post-harvest inulin degradation in *Cichorium intybus* than copy number variation of its close paralog 1-FEH IIa. *Frontiers in Plant Science*, 6, 455.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, 6(5), e19379.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, 14(8), 2611-2620.

Excoffier, L., Laval, G., & Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary bioinformatics online*, 1, 47.

Ersoz, E. S., Yu, J., & Buckler, E. S. (2007). Applications of linkage disequilibrium and association mapping in crop plants. In *Genomics-assisted crop improvement* (97-119). Springer Netherlands.

Flint-Garcia, S. A., Thornsberry, J. M., & IV, B. (2003). Structure of linkage disequilibrium in plants. *Annual review of plant biology*, 54(1), 357-374.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

Gibson, G.R., McCartney, A.L., & Rastall, R.A. (2006). Prebiotics and resistance to gastrointestinal infections. *British Journal of Nutrition* 93, S31-S34

Gouesnard, B., Bataillon, T. M., Decoux, G., Rozale, C., Schoen, D. J., & David, J. L. (2001). MSTRAT: An algorithm for building germ plasm core collections

- by maximizing allelic or phenotypic richness. *Journal of Heredity*, 92(1), 93-94.
- Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6), 226-231.
- Huynh, B. L., Mather, D. E., Schreiber, A. W., Toubia, J., Baumann, U., Shoaie, Z., ... & Fleury, D. (2012). Clusters of genes encoding fructan biosynthesizing enzymes in wheat and barley. *Plant molecular biology*, 80(3), 299-314.
- Itaya, N. M., Asega, A. F., Carvalho, M. A. M., & Rita de Cássia, L. (2007). Hydrolase and fructosyltransferase activities implicated in the accumulation of different chain size fructans in three Asteraceae species. *Plant Physiology and Biochemistry*, 45(9), 647-656.
- Jamann, T. M., Balint-Kurti, P. J., & Holland, J. B. (2015). QTL Mapping Using High-Throughput Sequencing. *Plant Functional Genomics: Methods and Protocols*, 257-285.
- Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403-1405.
- Karimi, R., Azizi, M.H., Ghasemlou, M., and Vaziri, M. (2015). Application of inulin in cheese as prebiotic, fat replacer and texturizer: a review. *Carbohydrate Polymer* 119, 85-100.
- Keenan, D.F., Resconi, V.C., Kerry, J.P., and Hamill, R.M. (2014). Modelling the influence of inulin as a fat substitute in comminuted meat products on their physico-chemical characteristics and eating quality using a mixture design approach. *Meat Science* 96, 1384-1394.
- Koopman, W., (2002). Zooming in on the lettuce genome: species relationships in *Lactuca s.l.*, inferred from chromosomal and molecular characters. *PhD Thesis*, Wageningen University, Netherlands

- Krutovsky, K. V., & Neale, D. B. (2005). Nucleotide diversity and linkage disequilibrium in cold-hardiness-and wood quality-related candidate genes in Douglas fir. *Genetics*, 171(4), 2029-2041.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14), 1754-1760.
- Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., ... & Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 28(18), 2397-2399.
- Li, Y., Haseneyer, G., Schön, C. C., Ankerst, D., Korzun, V., Wilde, P., & Bauer, E. (2011). High levels of nucleotide diversity and fast decline of linkage disequilibrium in rye (*Secale cereale* L.) genes involved in frost response. *BMC plant biology*, 11(1), 6.
- Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., & Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that corrects the bias due to population structure and relatedness. *Heredity*, 108(3), 285-291.
- Maroufi, A., Van Bockstaele, E., & De Loose, M. (2010). Validation of reference genes for gene expression analysis in chicory (*Cichorium intybus*) using quantitative real-time PCR. *BMC Molecular Biology*, 11(1), 15.
- Marita, J. M., Rodriguez, J. M., & Nienhuis, J. (2000). Development of an algorithm identifying maximally diverse core collections. *Genetic Resources and Crop Evolution*, 47(5), 515-526.
- Maroufi, A., Van Bockstaele, E. & De Loose M.. 2012. “Differential Expression of Fructan 1-Exohydrolase Genes Involved in Inulin Biodegradation in Chicory (*Cichorium Intybus*) Cultivars.” *Australian Journal of Crop Science* 6(9): 1362–68.
- Mendoza, E., Garcia, M.L., Casas, C., & Selgas, M.D. (2001). Inulin as fat

substitute in low fat, dry fermented sausages. *Meat Science* 57, 387-393.

Meyer, D., & Stasse-Wolthuis, M. (2009). The bifidogenic effect of inulin and oligofructose and its consequences for gut health. *European Journal of Clinical Nutrition* 63, 1277-1289

Müller, J., Aeschbacher, R. A., Sprenger, N., Boller, T., & Wiemken, A. (2000). Disaccharide-mediated regulation of sucrose: fructan-6-fructosyltransferase, a key enzyme of fructan synthesis in barley leaves. *Plant Physiology*, 123(1), 265-274.

Muys, C., Thienpont, C. N., Dauchot, N., Maudoux, O., Draye, X., & Cutsem, P. V. (2014). Integration of AFLPs, SSRs and SNPs markers into a new genetic map of industrial chicory (*Cichorium intybus* L. var. sativum). *Plant breeding*, 133(1), 130-137.

Ohta, T., & Kimura, M. (1975). The effect of selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect). *Genetical research*, 25(03), 313-325.

Pembleton, L. W., Wang, J., Cogan, N. O. I., Pryce, J. E., Ye, G., Bandaranayake, C. K., ... & Forster, J. W. (2013). Candidate gene-based association genetics analysis of herbage quality traits in perennial ryegrass (*Lolium perenne* L.). *Crop and Pasture Science*, 64(3), 244-253.

Prasad, A., Schnabel, R. D., McKay, S. D., Murdoch, B., Stothard, P., Kolbehdari, D., ... & Moore, S. S. (2008). Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle. *Animal genetics*, 39(6), 597-605.

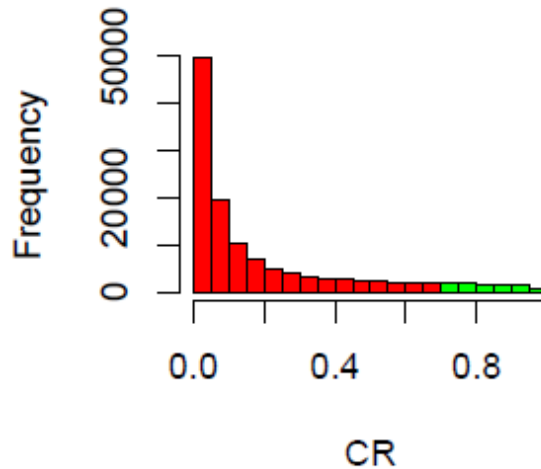
Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945-959.

- Raulier, P., Maudoux, O., Notté, C., Draye, X., & Bertin, P. (2015). Exploration of genetic diversity within *Cichorium endivia* and *Cichorium intybus* with focus on the gene pool of industrial chicory. *Genetic Resources and Crop Evolution*, 1-17.
- Ruttink, T., Veeckman, E., Haegeman, A., van Parijs, F., Van Glabeke, S., Muylle, H., & Roldán-Ruiz, I. (2015). Probing genome diversity in natural and breeding populations of *Lolium perenne*. int. *Plant Animal Genome Conference*, PAG XXIII, San Diego, United States.
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112-118.
- Tamura, K. I., Kawakami, A., Sanada, Y., Tase, K., Komatsu, T., & Yoshida, M. (2009). Cloning and functional analysis of a fructosyltransferase cDNA for synthesis of highly polymerized levans in timothy (*Phleum pratense* L.). *Journal of experimental botany*, 60(3), 893-905.
- Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., & Gaut, B. S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proceedings of the National Academy of Sciences*, 98(16), 9161-9166.
- van Arkel J, Vergauwen R, Sévenier R, Hakkert JC , van Laere A, Bouwmeester H J, ... & van der Meer IM (2012) Sink Filling, Inulin Metabolizing Enzymes and Carbohydrate Status in Field Grown Chicory (*Cichorium Intybus* L.).” *Journal of Plant Physiology* 169:1520–29.
- van Arkel J , Sévenier R , Hakkert J C , Bouwmeester, H. J., Koops, A. J., & van der Meer, I. M. (2013). Tailor-made fructan synthesis in plants: a review. *Carbohydrate polymers*, 93(1), 48-56.

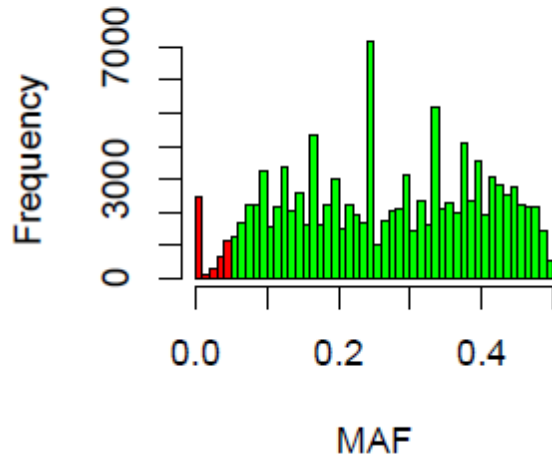
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11), 4414-4423.
- Van Waes, C., Baert, J., Carlier, L., & Van Bockstaele, E. (1998). A rapid determination of the total sugar content and the average inulin chain length in roots of chicory (*Cichorium intybus* L). *Journal of the Science of Food and Agriculture*, 76(1), 107-110.
- Vergauwen, R., Van Laere, A., & Van den Ende, W. (2003). Properties of fructan: fructan 1-fructosyltransferases from chicory and globe thistle, two Asteracean plants storing greatly different types of inulin. *Plant physiology*, 133(1), 391-401.
- Vijn, I., & Smeekeens, S. (1999). Fructan: more than a reserve carbohydrate?. *Plant physiology*, 120(2), 351-360.
- Wang, C., Van den Ende, W., & Tillberg, J. E. (2000). Fructan accumulation induced by nitrogen deficiency in barley leaves correlates with the level of sucrose: fructan 6-fructosyltransferase mRNA. *Planta*, 211(5), 701-707.
- Wang, C., Van den Ende, W., & Tillberg, J. E. (2000). Fructan accumulation induced by nitrogen deficiency in barley leaves correlates with the level of sucrose: fructan 6-fructosyltransferase mRNA. *Planta*, 211(5), 701-707.
- Wilson, R. G., Smith, J. A., & Yonts, C. D. (2004). Chicory root yield and carbohydrate composition is influenced by cultivar selection, planting, and harvest date. *Crop Science*, 44(3), 748-752.
- Zhang, J., Huang, S., Fosu-Nyarko, J., Dell, B., McNeil, M., Waters, L., ... & Appels, R. (2008a). The genome structure of the 1-FEH genes in wheat (*Triticum aestivum* L.): new markers to track stem carbohydrates and grain filling QTLs in breeding. *Molecular Breeding*, 22(3), 339-351.

- Xing, Y., Frei, U., Schejbel, B., Asp, T., & Lübberstedt, T. (2007). Nucleotide diversity and linkage disequilibrium in 11 expressed resistance candidate genes in *Lolium perenne*. *BMC plant biology*, 7(1), 43.
- Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., ... & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4), 355-360.
- Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5), 614-620.

8.8 Supplementary materials



Supplementary material 16 :Histogram of calling rate (CR) of the 120 000 SNPs identified by GBS. Green, SNPs with a CR value above threshold of 0.66 .



Supplementary material 17 Histogram of minor allele frequency (MAF) of the 120 000 SNPs identified by GBS. Green, SNPs with a MAF value above threshold of 0.05.

			Harvest							
			Fresh weight	DMC	CC	IN	DPin	FGn	FFn	SUCn
Harvest	DMC	-0,36								
	CC	-0,16	0,56							
	IN	-0,20	0,62	0,50						
	DPin	-0,21	0,29	0,01	0,57					
	FGn	0,17	-0,58	-0,50	-0,61	-0,23				
	FFn	0,11	-0,51	-0,39	-0,53	-0,14	0,67			
	SUCn	0,19	-0,55	-0,45	-0,98	-0,57	0,45	0,35		

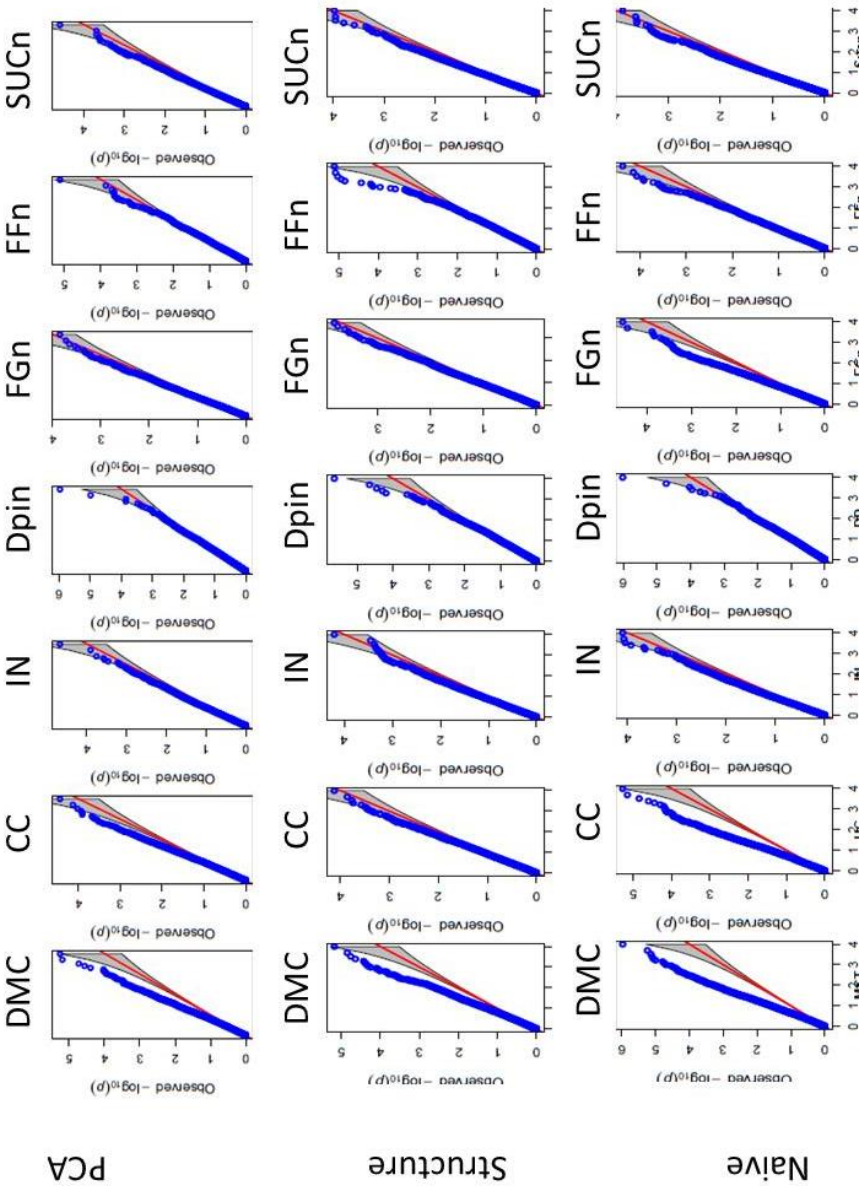
Supplementary material 18: Correlation between the fresh weight (FW) dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), Free glucose content (FGn), Free fructose content (FFn) and sucrose content (SUCn) measured at harvest.

		Harvest							
		Fresh weight	DMC	CC	IN	DPin	FGn	FFn	SUCn
Variation after cold exposure	DMC	<u>0,01</u>	0,19	0,11	0,02	-0,01	-0,12	-0,11	0,01
	CC	-0,14	0,13	0,28	0,11	0,03	-0,10	-0,10	-0,10
	IN	0,00	-0,27	-0,10	-0,28	-0,15	0,13	0,03	0,30
	DPin	-0,20	0,17	0,02	0,22	0,34	-0,15	-0,17	-0,20
	FGn	-0,09	0,18	0,02	0,14	0,05	0,03	0,05	-0,17
	FFn	-0,13	0,41	0,18	0,36	0,23	-0,28	-0,15	-0,35
	SUCn	0,06	0,16	0,05	0,19	0,10	-0,04	0,04	-0,22

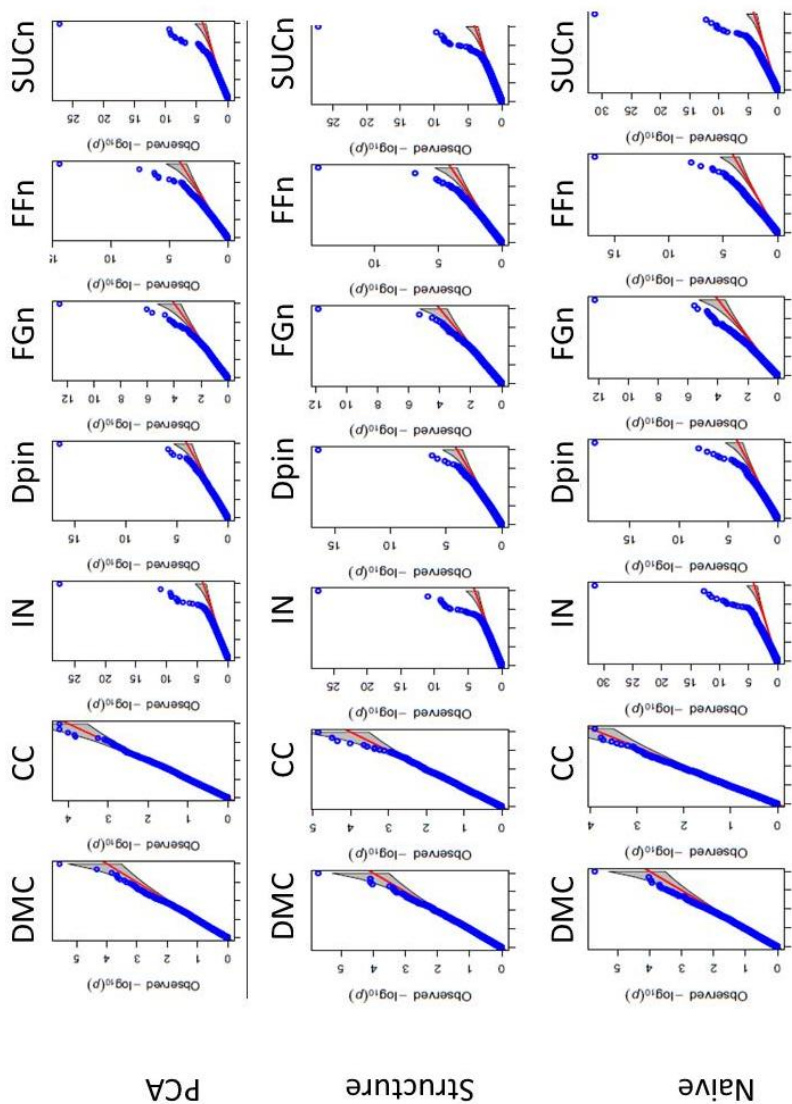
Supplementary material 19: Correlation between the fresh weight (FW) dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), Free glucose content (FGn), Free fructose content (FFn) and sucrose content (SUCn) measured at harvest and the variation after the cold exposure.

		Variation after cold exposure							
		Fresh weight	DMC	CC	IN	DPin	FGn	FFn	SUCn
Variation after cold exposure	DMC								
	CC		-0,09						
	IN		0,28	-0,11					
	DPin		0,25	-0,20	0,54				
	FGn		-0,36	-0,05	-0,68	-0,24			
	FFn		-0,21	0,06	-0,76	-0,04	0,68		
	SUCn		-0,26	0,11	-0,95	-0,51	0,56	0,53	

Supplementary material 20:Correlation between variation, after a cold exposure of the fresh weight (FW) dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), Free glucose content (FGn), Free fructose content (FFn) and sucrose content (SUCn).



Supplementary material 21: QQ plot of the p-values obtained with the PCA model, the Structure model and the naive model for the fresh weight (FW) dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), Free glucose content (FGn), Free fructose content (FFn) and sucrose content (SUCn) measured at harvest.



Supplementary material 22 : QQ plot of the p-values obtained with the PCA model, the Structure model and the naive model for the variation after cold exposure in fresh weight (FW) dry matter content (DMC), carbohydrate content (CC), inulin content (IN), mean polymerization degree (DPin), Free glucose content (FGn), Free fructose content (FFn) and sucrose.

9 Discussion and conclusion

This thesis aimed to unravel the genetic determinism of carbohydrate (including inulin) synthesis and the cold-induced inulin degradation kinetics occurring in field. The three specific objectives were: (i) the characterization of the genetic variability in the *Cichorium intybus* germplasm, (ii) the confirmation of the presence of major effect QTLs accounting for the cold-induced inulin degradation –i.e. the candidate gene approach- and (iii) the identification of minor effect QTLs associated to carbohydrate (including inulin) quality at harvest and cold-induced inulin degradation –i.e. the genome wide association study-. These three specific objectives were addressed in seven work packages. Here, we will discuss our main results, point to the advances of the thesis and the shortcomings of our approach, considering each work package.

9.1 Work package 1: Characterization of the genetic diversity of industrial chicory

The investigation of the genetic diversity of the *Cichorium intybus* germplasm validated and improved the current classification and our understanding of the genetic structure of the wild and cultivated germplasms. The proximity between some wild *Cichorium intybus* and the *Cichorium endivia* highlighted the complex breeding history of *C. endivia* and provided new clues for future studies aiming at identifying the origin of *C. endivia*.

We confirmed the subdivision of the cultivated germplasm into three groups: the leaf chicories, the Witloof and the root chicories. Among the leaf chicories, we validated the differentiation between the Radicchio and Sugarloaf cultivar types and identified a new group which we named “Catalogne”. We hope this finding will decrease the confusion concerning the identification of this group, regularly considered as wild chicory.

Apart from the Witloof, all cultivated cultivar groups and sub-groups presented a high level of genetic diversity, revealing —although not demonstrating— a potential gene pool for the breeding purpose, i.e. as a resource of genes associated to interesting traits such as yield, quality, disease resistance, etc. Besides, the small linkage disequilibrium (LD) in modern industrial cultivars suggests that a small LD was already present within the ancient root cultivars such as Magdeburg and within the wild populations. Indeed, the LD extend tend to be larger in populations that experiment bottleneck events or intensive selection, such as elite breeding material (Breseghello and Sorrels 2006, Flint-Garcia et al. 2003).

9.2 Work package 2: Creation of a diversity panel suitable for candidate-gene and genome-wide studies

The diversity panel needed for genome-wide association analysis was established from the cultivars at the origin of the breeding of the current cultivars. It is thus (i) expected to embrace the genetic diversity of the current cultivars, (ii) to include alleles associated to interesting characters that were lost during the past 40 years of breeding, and finally (iii) this panel is phenotypically close to current cultivars (i.e. root shape), so that phenotyping and introgression of interesting alleles would be possible. The sampling method appeared suitable to generate a panel representing the collection analyzed. This method has also been shown to conserve the population structure (Jamann et al. 2015).

The low LD of a panel is synonymous for a high mapping accuracy of the causative variants provided that a high density of markers is available. This points to the main limitations of our work that could not be circumvented as the LD of the panel was estimated after its creation. (i) For the candidate approach, the markers density in candidate genes may not be sufficient to detect possible causative variants present in other regions of the gene. (ii) For the GWAS, the marker density is not sufficient to explore the whole genome.

The characterization of our panel enables us to better anticipate the creation of new panels for future projects in accordance with the aims of the studies and the financial resources. For studies aiming at identifying the genetic basis of qualitative traits such as disease resistances, the literature advises panels presenting high level of genetic diversity (Rikkerink et al. 2007). In this case, a panel including ancient “coffee” cultivars, others cultivar groups or wild individuals could be interesting if appropriate genotyping technologies were available to cope with low LD. Nevertheless, this type of panel would not be suitable for quantitative traits related to the carbohydrate profile or yield. Indeed, the elongated root of ancient cultivars or the thin or branched root of leaf cultivars and wild individuals would drastically influence (i) the ability to harvest the roots, (ii) the ability to phenotype the carbohydrate profile and finally, (iii) the physiological components underlying the traits. In this context, a panel similar to the one created in this study or a panel including elite lines would be better suited. The latter would present a reduced genetic diversity and an increased LD. Since the LD is inversely proportional to the genetic diversity, whatever the purpose of the study, there would be a trade-off between genetic diversity (i.e. a panel including more genetic diversity potentially includes more valuable alleles) and the LD (i.e. a panel with lower LD results in a higher accuracy of the QTL localization, yet increases the genotyping effort).

9.3 Work package 3: Phenotyping

Three field trials were conducted to quantify the carbohydrate synthesis and inulin degradation under autumnal conditions for all the individuals of the

diversity panel. The autumnal conditions were simulated by exposing the roots to cold temperatures, and the degradation was inferred from one measurement at harvest and one after a storage period. Since the carbohydrate analysis is destructive and the individuals of the panel cannot be cloned, we implemented a progeny-test strategy that consisted in selfing the individuals of the panel and phenotyping their progeny during the next season. To our knowledge, this is the largest phenotyping effort in fructan species.

The phenotyping led to the identification of some lines presenting some superior quality standards sought for the elite lines (shape, yield or fructan profile), revealing the usefulness of the ancient germplasm to improve the breeding material.

The first limitation of our phenotyping strategy was the repartition of the phenotyping on three different trials and the absence of repetition for most lines (three lines and two current cultivars could be repeated three times in the three trials). The inclusion of all lines in one trial and the increase of repeats within and across the trials would have improved the accuracy of the phenotypic evaluation and the evaluation of the trial effect. Unfortunately, due to practical and financial constraints, it was not possible to carry the three trials during the same season.

The second limitation of our phenotyping was the change of carbohydrate analysis method across the different trials. HPLC was used for the first trial, and the DP was evaluated as the ratio between the number of fructose and glucose units trapped in fructans. In presence of inulose, this DP value was

overestimated. During the second and third trials, HPLC and HPAEC analysis were combined to generate the complete fructan profile and estimate the true DP of both inulin and inulose.

The trials yielded the opportunity to phenotype others traits: the root shape by means of image analysis and the complete fructan profile that describes the content of inulin and inulose of each DP. According to available resources, the analysis of these traits was completed during the thesis. In particular, methods to analyze the fructan profile and the root shape were developed during two master thesis (Nicolas Urbin-Choffray 2012 and Nicolas Saintmard 2015). A genome-wide association mapping study concerning the root shape was also realized during the master thesis of Nicolas Saintmard.

9.4 Work package 5: Genotyping

The genome-wide genotyping of the panel has been a major component of the project. At the origin of the project, the planning was to generate 200 markers by EcoTILLING of an F2 population and to investigate the polymorphisms present in EST bank. Thanks to the sequencing advances, we could use the Genotyping-by-Sequencing (GBS) technique to genotype the whole panel with a large number of markers, high enough (9 859 SNPs) for GWAS. We also identified many other polymorphic loci (112 791 SNPs) that may be used to generate new markers in regions of interest such as QTLs or candidates genes of

many traits, and especially around the significant SNPs identified in our GWAS study.

Nevertheless, the GBS technology could not completely circumvent the problem of the high marker density required for the GWAS. The number of markers could be increased by extending the sequencing of the available GBS libraries or initiating new GBS using other restriction enzymes such as PST I and Taq I. As the cost of sequencing is continuously decreasing, this should be accessible in a reasonable time frame.

To anchor the GBS SNPs on a physical map, we initiated the sequencing and assembly of a reference genome. At the time of writing, the reference genome is still an incomplete puzzle. It is currently being assembled in larger pieces via a recent collaboration with UGhent. This new genome assembly will be an outstanding tool for breeders and researchers working on chicory and related species.

9.5 Work packages 4 & 6: identification and validation of candidate genes

Initially, Nicolas Dauchot identified polymorphisms in candidate genes by EcoTILLING on two groups of experimental lines with contrasted inulin degradation after cold exposure (depolymerisation sensitive vs depolymerisation resistant). These lines had been obtained after few

selection cycles from an initial population. This strategy has proven powerful to detect the major QTL responsible for these contrasting phenotypes, the QTL including the three FEH.

The association between the polymorphisms in the three FEH genes and the intensity to inulin degradation was confirmed by the candidate-gene approach and by the GWAS. These analyses, combined to the genome analysis, revealed that the three FEH genes are located in one genomic region (at least 360 Kb) that presents a very high LD compared to the genome-wide LD. A detailed analysis (published in Dauchot et al. 2015) revealed that the deletion in an active site of *1-FEH IIb* is responsible for the reduction of inulin degradation. Individuals carrying this deletion present a reduced 1-FEH activity, such as presented in the third ideotype presented in the introduction (Figure 51). The remaining 1-FEH activity in individuals carrying the deletion implies that other 1-FEHs are still active and therefore other QTLs may still be identified. As no other major QTLs have been identified and since the heritability of sensitivity to inulin depolymerisation is low, we expect that part of the remaining genotypic variance is the result of QTLs with minor effect such as the one identified in the GWAS analysis.

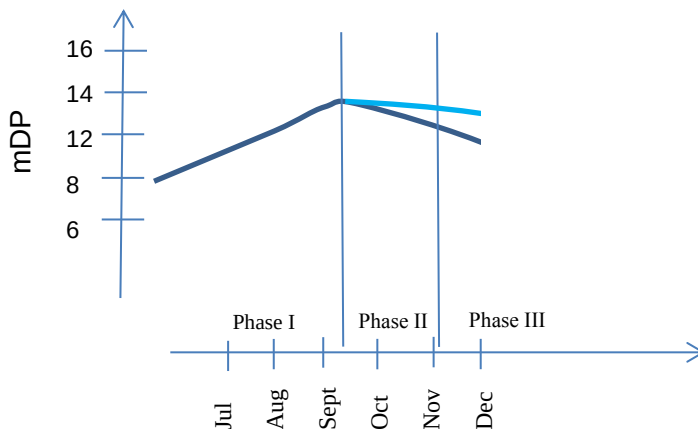


Figure 51: Change in inulin mDP in normal chicory taproot (dark blue, inspired from a figure from van Arkel et al. 2012.) and in the chicory ideotype with a low 1-FEH activity.

The *1-FEH IIb* deletion was present in most cultivars at the origin of the current cultivars. Since it has a major effect on the susceptibility to inulin depolymerisation, it has been rapidly fixed in the elite material of Chicoline (C. Notté and O. Maudoux, personal communication). The use of the deletion as a marker for marker assisted selection is discussed later.

The candidate-gene approach also investigated the polymorphisms within other genes of the GH32 family, including the fructan-synthesis genes *1-SST*, *1-FFT* and some invertase genes involved in sucrose degradation. These polymorphisms were not associated to the phenotypic variation of any trait describing the carbohydrate and inulin quality. In the introduction, we proposed that the variation in enzyme affinity for long or short fructan could be at the origin of the phenotypic variability in fructan profiles (Figure 52). Nevertheless, the negative

results concerning *I-SST*, *I-FFT* and the invertases suggests that, in the industrial chicory germplasm, the chance of identifying polymorphism in fructan genes associated to a change in fructan enzyme affinity is very low.

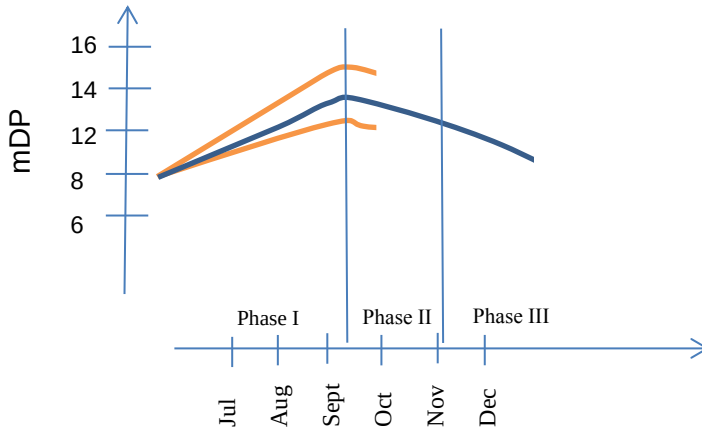


Figure 52: Change in inulin mDP in normal chicory taproot (dark blue, inspired from a figure from van Arkel et al. 2012.) and the chicory ideotype characterized by high or low 1-FFT affinity for low DP inulin (orange).

9.6 Work package 7: Genome wide investigation of genetic factors associated to variation on carbohydrate synthesis and cold-induced inulin degradation.

The traits associated to the inulin synthesis and degradation showed low heritabilities. Accordingly, and apart from the *I-FEH IIb* QTL, only minor effect SNPs were identified for fresh weight, dry matter content, inulin concentration, polymerization degree of fructose content and finally, inulin degradation. These findings are consistent with the fact that these traits are the outcome of several inter-related physiological processes. The underlying genes should therefore be pleiotropic, in line with the observation that QTLs were not located in genes of the fructan pathway. These mechanisms could be related (i) to sink-source relations, (ii) to variation of traits that affect the expression of fructan genes (Figure 53) as observed in other species.

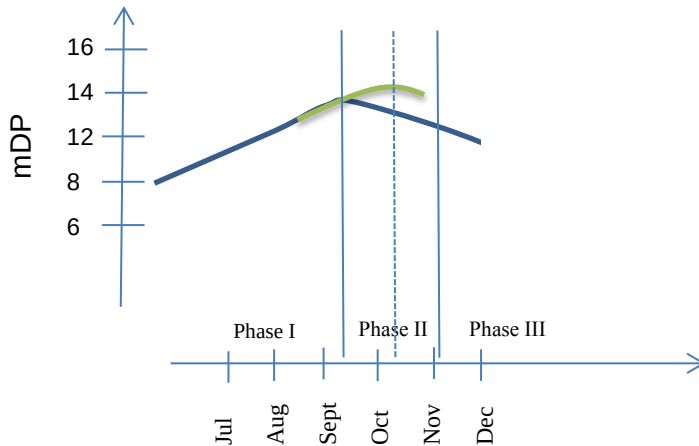


Figure 53: Change in inulin mDP in normal chicory taproot (dark blue, inspired from a figure from van Arkel et al. 2012.) and the chicory ideotype presenting an extended duration of 1-SST expression (green).

Different directions can be considered to improve the quality and the usefulness of the GWAS results for marker-assisted selection in chicory. These are discussed in the following paragraphs.

Firstly, the improved assembly of the genome will streamline the mapping of the different markers. In particular, when several SNPs were associated to one trait, the assembly will clarify if these SNPs correspond to one or several QTLs. In a further step, the gene prediction on the genome assembly will help to identify candidate-genes in the vicinity of significant SNPs.

Secondly, using the genome assembly, it will be possible to increase the marker density around the significant SNPs. This would result in two different

improvements. Firstly, it could validate the identified QTL. Secondly, it could increase the accuracy of the QTL mapping, potentially elucidate the underlying physiological mechanism and could increase the accuracy of the evaluation of the proportion of phenotypic variance explained by the causative polymorphism. The SNPs identified by GBS could be used to increase the local marker density (see above). If these SNPs are not sufficient, new SNPs can be targeted in the region of interest.

Thirdly, we would increase the genome wide SNP density in order to identify new other remaining QTLs. We already discussed earlier how GBS and the decreasing sequencing cost could help to increase the genome coverage in a close future.

Fourthly, the marker assisted selection can either be used in a breeding population to select the seedlings presenting an interesting allele, or to optimize the introgression of an interesting allele in an elite genotype. The strategy implemented will depend of the occurrence of the interesting allele in the breeding material. Three different scenarios are expected:

- 1) The interesting allele may be **fixed** in the breeding material, which is the case of the 1-FEH I1b deletion. Such a marker has no direct interest for the improvement of the current breeding material. Nevertheless it may be very useful for the introduction of new individuals -such as the superior individuals identified in the diversity panel- that do not carry the interesting allele. The breeder can select the progenies of the interesting individual and elite material, based on this marker.

- 2) The interesting allele may be present in **intermediate frequency** in the breeding material. In that situation, the breeder may use the marker to increase the frequency of the interesting allele in the breeding material. In such situation the breeder may investigate marker assisted strategies, such as gene pyramiding, a strategy aiming at gathering all QTLs in one individual.
- 3) The interesting allele may be **absent** from the breeding material. Therefore, the interesting allele may be introgressed in an individual presenting a general superior phenotype by recurrent back-crossing. Recurrent back-crossing is a strategy aiming at creating an individual genetically similar to a “recurrent” parent but also including one or several traits of a “donor” parent. The method consists in crossing the donor and recurrent parent and then crossing their progenies with the recurrent parent, and this several time. The number of back-cross may be reduced using markers covering different genomic sections in order to select individuals carrying most of the of genome of the recurrent parent and the gene(s) associate to the interesting trait of the donor parent.

9.7 General conclusion

In conclusion, we concluded that the genetic determinism inulin and carbohydrates synthesis is multigenic and determined by several minor effect QTLs when cold induced inulin degradation result from the action of one major effect QTL and at least one minor effect QTLs.

9.8 Reference

- Bresegghello, F., & Sorrells, M. E. (2006). Association analysis as a strategy for improvement of quantitative traits in plants. *Crop Science*, 46(3), 1323-1330.
- Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. (2003). Structure of linkage disequilibrium in plants*. *Annual review of plant biology*, 54(1), 357-374.
- Jamann, T. M., Balint-Kurti, P. J., & Holland, J. B. (2015). QTL Mapping Using High-Throughput Sequencing. *Plant Functional Genomics: Methods and Protocols*, 257-285
- Rikkerink, E. H., Gardiner, S. E., & De Silva, H. N. (2007). *Association mapping in plants*. New York: Springer.