



Facultés des Sciences
Département de Biologie
Unité de Recherche en Biologie Cellulaire et Moléculaire Végétale

Cartographie génétique de *Cichorium intybus* L. : Marqueurs moléculaires et synténie

Dissertation présentée par
Céline Muys
en vue de l'obtention du
grade de Docteur en Sciences

Composition du jury :

Dr Nicolas Dauchot (FUNDP - Namur)
Pr. Xavier Draye (UCL - Louvain-la-Neuve)
Dr. Olivier Maudoux (Cosucra Groupe Warcoing SA)
Pr. Johan Messiaen (FUNDP - Namur)
Pr. Pierre Van Cutsem (promoteur) (FUNDP - Namur)

2012

Cartographie génétique de *Cichorium intybus* L. : marqueurs moléculaires et synténie

Par Céline Muys

La Belgique est le premier producteur mondial de chicorée industrielle. Cette espèce est cultivée afin d'en extraire l'inuline, un polysaccharide de réserve utilisé dans l'agro-alimentaire et aux vertus pré-biotiques. La sélection des variétés cultivées de chicorée est majoritairement phénotypique et basée essentiellement sur le rendement en sucre, la montaison, la résistance aux maladies ainsi que sur la morphologie racinaire. L'objectif de ce travail consiste à développer des outils moléculaires pouvant améliorer et faciliter cette sélection.

Les banques EST (Expressed Sequence Tag) ont été utilisées afin de développer des marqueurs géniques de type SSR (Simple Sequence Repeats) et SNP (Single Nucleotide Polymorphism). Au total, 130 marqueurs ont été validés chez la chicorée industrielle. Ces marqueurs ont été utilisés pour la construction d'une carte génétique de 1208 cM composée de 237 marqueurs (152 marqueurs AFLP, 27 marqueurs SNP et 40 marqueurs SSR) en collaboration avec l'UCL et la société Chicoline. Les séquences transcrites associées aux marqueurs géniques cartographiés ont ensuite été utilisées pour la recherche d'orthologues chez la laitue afin d'élaborer une carte comparative entre les deux espèces. Cette analyse a permis de mettre en évidence des régions de synténie dans lesquelles se trouvaient des gènes candidats positionnels potentiellement impliqués dans divers mécanismes d'intérêt pour la sélection de la chicorée.

Cichorium intybus L. genetic map: molecular markers and synteny

By Céline Muys

Belgium is the world's largest producer of industrial chicory. This species is cultivated to extract inulin, a storage polysaccharide with pre-biotic properties used in food processing. The selection of chicory varieties is mainly phenotypic and based essentially on sugar yield, bolting, disease resistance and root morphology. The objective of this work is to develop molecular tools that can improve and facilitate breeding.

EST (Expressed Sequence Tag) libraries were used to develop SSR (Simple Sequence Repeats) and SNP (Single Nucleotide Polymorphism) genic markers. A total of 130 markers were validated for industrial chicory. These markers were used to build a genetic map of 1208 cM composed of 237 markers (152 AFLP markers, 27 SNP markers and 40 SSR markers) in collaboration with UCL and the Chicoline Company. The transcribed sequences associated with mapped genic markers were then used to search for orthologs in lettuce to develop a comparative map between the two species. This analysis has allowed identifying regions of synteny in which there were positional candidate genes potentially involved in different mechanisms of interest for the selection of chicory.

« *Les vraies passions sont égoïstes* »

Stendhal; Le rouge et le noir

Remerciements...

Sept années auront été nécessaires pour clôturer ce travail. Sept années pendant lesquelles je me suis épanouie dans ma vie professionnelle et dans ma vie privée afin de me construire et de devenir la scientifique, la femme et la mère que je suis aujourd'hui... Tout cela n'aurait pas été possible sans l'aide, le soutien ou simplement la présence de nombreuses personnes que je tiens à remercier dans ces quelques lignes.

Merci au professeur Pierre Van Cutsem et au Professeur Johan Messiaen en tant que directeurs de l'URBV pour m'avoir accueilli au sein de leur laboratoire il y a 9 ans déjà.

Merci encore au Professeur Pierre Van Cutsem pour son encadrement en tant que Promoteur. Merci pour les nombreuses corrections, remarques et relectures concernant ce manuscrit. Merci de m'avoir fait confiance, il y a 7 ans en m'engageant comme assistante. J'ai adoré ce travail que ce soit au niveau didactique ou scientifique et c'est le cœur lourd que je clôture ma mission dans votre équipe.

Merci aux membres du jury, le Professeur Xavier Draye, Olivier Maudoux, le professeur Johan Messiaen, et Nicolas Dauchot pour le temps qu'ils ont consacré à la lecture de cette thèse ainsi que pour les remarques constructives qui ont permis de l'améliorer.

Un tout grand merci à Claire-Noëlle Thienpont sans qui ce travail n'aurait pas été possible. Cette collaboration fut pour moi très enrichissante. Merci de m'avoir éclairé sur de nombreux points concernant la cartographie. Ton travail et tes conseils m'ont été très précieux. Bonne route...

Merci à Lionel Leclerc pour sa contribution lors de l'analyse des microsatellites.

Merci à Richard Mairesse de m'avoir éclairé dans les méandres du monde Linux.

Je suis arrivée au laboratoire en 2004 afin d'y réaliser mon mémoire. Déjà, à cette époque, Nicolas, tu as accepté de m'encadrer et de me guider dans la formidable expérience des marqueurs moléculaires. Un tout grand merci pour tout ce que tu m'as appris, pour tes conseils, tes analyses, tes éclaircissements, tes innombrables bonnes idées et de toujours avoir pris le temps de répondre à mes très très nombreuses questions.

Merci à toutes les personnes que j'ai appris à connaître au sein de l'URBV et qui sont pour moi comme une seconde famille. Merci à Françoise pour sa grande disponibilité, son écoute, sa serviabilité. Merci de m'avoir facilité la tâche pour les aspects administratif et technique des TP. Merci à Pierre d'avoir une solution à chaque problème. Merci également pour les nombreuses sorties mémorables que nous avons faites ensemble. J'espère qu'il y en aura encore beaucoup d'autres. Merci à Aurélien pour nos nombreuses discussions. Ce fut un plaisir de t'avoir à mes côtés (au sens propre comme au figuré). Les jeudis sont bien vides depuis que tu es parti.... Merci à Catherine, Géraldine, Geoffroy, Adrien, François, Roland, Rafael, Houda ainsi qu'à tous ceux qui ont travaillé dans ce laboratoire. Merci d'avoir fait de l'URBV un endroit où il y fait bon vivre.

Merci à Vincent, Bénédicte et Véronique pour les TP donnés ensemble. Un merci tout particulier à Kayleen avec qui j'ai passé d'excellents moments. Merci pour toutes ces discussions, ces fous rires durant ces centaines d'heures de TP à donner, redonner et re-redonner les mêmes explications. Notre équipe de choc risque de beaucoup me manquer. Bonne continuation et bonne chance pour l'année prochaine. Merci également à Xavier, Aurélie et Sophie pour les bons moments passés lors des deux semaines annuelles de TP intégrés.

Petit clin d'œil à Isa, mon amie de très longue date qui m'a guidé vers la Biologie en fin de retho. Merci pour ces innombrables soirées passées ensemble, pour ces marathons sans queue ni tête, pour ces fous rires et pour tout ce qui fait notre amitié...

La biologie ne m'aurait jamais parue si intéressante sans la rencontre il y a 10 ans de cela, de mes sœurs d'armes : Audrey, Audrey, Hélène et Vanessa. Merci à vous quatre de jouer un rôle important dans ma vie. Il y a pour chacune d'entre vous une place spéciale dans mon cœur.

Merci à mon ami Jojo d'avoir toujours été présent dans les moments de fête mais aussi et surtout dans les moments plus difficiles. Je sais au plus profond de moi que nous pourrons toujours compter l'un sur l'autre. Merci à Guillaume, Thomas et Gaby de m'avoir toujours accepté comme je suis réellement.

Merci à Véro, Sophie, Méla, Sophie, Céline, Isa, Hélène, Julie, Pupu, Babeth, Kikine, Amé, Delphine, Briec, Jeff, Sarah, Charline de former un groupe d'amis formidables avec lequel c'est un vrai plaisir de discuter, de faire la fête et de décompresser, surtout lors de ces derniers mois difficiles.

Un tout grand merci à ma très chère grande sœur qui est toujours très attentive à mes besoins. Merci pour tes très nombreuses intentions. Merci de m'avoir sauvé de la déshydratation et de la faim dans les moments les plus difficiles de ma rédaction. Merci pour toutes ces petites choses que tu as faites et qui ont rendu ce travail un peu moins difficile.

Merci à mon cher petit frère adoré pour sa présence, pour les baby-sittings, pour les coups de main. Merci aussi de répondre presque toujours au téléphone et de le recharger à temps ; c'est un vrai soulagement !

Merci à mes parents. Je me permets de vous citer ensemble car quoi qu'on en dise, c'est le couple que je remercie. Merci d'avoir fait de moi une battante, une fonceuse et une passionnée. Sans vous deux, il est certain que je n'en serai pas là aujourd'hui. Merci, merci, merci...

Je voudrai aussi vous remercier séparément. Merci maman pour les innombrables corrections orthographiques de ce manuscrit. Merci à toi et à Marc d'être toujours aussi accueillants et prévenants vis-à-vis de nous et nos enfants. Merci à toi, Papa, de te soucier, à ta façon,

de notre bien-être et de toujours veiller à ce que chacun ne manque de rien. Zarah, Yassin, prenez à votre tour soin de lui.

Merci Mimi d'avoir été attentive à nos besoins ces derniers mois et d'avoir facilité notre vie quotidienne. Par ces innombrables gestes, vous avez contribué massivement à la réalisation de ce travail. Merci à Henry, de nous avoir prêté Mimi et toujours avec le sourire.

Merci à tous les autres membres de ma famille : Marraine l'autre, Papy, mes innombrables oncles, tantes, cousins, cousines, neveux, nièces et filleul. Merci à Charlotte, Marie-Pierre, Stéphane, Laurent, Véronique, Caroline, Virginie, Arnaud, Antoine, Mathieu. Merci d'avoir toujours pris le temps de venir aux nouvelles et de me remonter le moral dans les moments clés. Merci à Dédé pour la toute dernière correction orthographique.

Un immense merci à mes deux petits monstres, Lily-Rose et Gaspard. Qui sait, peut-être qu'un jour, vous trouverez ce manuscrit au fond d'une vieille armoire et que vous lirez ces quelques lignes. Alors merci de m'avoir appris, oh combien il était précieux, bénéfique et vivifiant de passer chaque minute disponible avec vous. Merci pour les sourires que vous m'avez offerts et qui m'ont permis de tenir le cap. Je vous aime de tout mon cœur.

Et enfin, last but not least, mon cher et tendre époux, Maxime. Cette thèse, c'est moi qui l'ai écrite mais c'est toi qui l'as épaulée. Comme tu l'appelles, Thèse Tassin, notre troisième enfant ne nous aura pas épargnés. Merci du fond du cœur de m'avoir supporté pendant ces longs mois de rédaction. Je sais, oh combien, ça a été difficile de supporter mes longues heures passées dans le bureau, de supporter mon stress quotidien, mes sauts d'humeur, mes besoins excentriques de décompression, mes airs absents. Merci d'avoir pris le relais dans la gestion des tâches quotidiennes et de t'être occupé des enfants quand la chiorée m'appelait. Tu es un mari formidable. Voilà mon cœur, c'est fini, je suis de retour... Merci, merci, merci.

Merci à tous ceux, qui, de près ou de loin, m'ont aidé pour ce travail. Merci à ceux, que j'ai oublié, de ne pas se vexer, ma mémoire me fait défaut depuis que je la consacre à la chiorée...<3

INTRODUCTION GÉNÉRALE	1
CHAPITRE 1. REVUE BIBLIOGRAPHIQUE	3
I. Cichorium intybus var. sativum : classification, intérêt et sélection	3
I.I. Classification taxonomique de la chicorée industrielle.....	3
I.II. Intérêt de la chicorée industrielle	4
I.II.1. Historique de la culture de chicorée et production	4
I.II.2. L'inuline : définition, synthèse, rôle et utilisation	5
I.III. La sélection chez la chicorée industrielle	10
II. Les marqueurs moléculaires	11
II.I. Introduction	11
II.II. Les microsatellites.....	13
II.II.1. Introduction	13
II.II.2. Mécanismes d'évolution et taux de mutation des microsatellites	14
II.II.3. Les microsatellites dans les régions transcrites	16
II.III. Les SNP (Single Nucleotide Polymorphism)	20
II.III.1. Introduction	20
II.III.2. Fréquence et influence des mutations dans les régions transcrites	22
II.IV. Les banques EST : sources de polymorphisme SSR et SNP.....	24
II.IV.1. Introduction	24
II.IV.2. Les banques EST chez les Asteraceae.....	24
II.IV.3. Avantages des marqueurs issus de banques EST	25
II.IV.1. Génotypage des marqueurs issus de banques EST	27
II.V. Méthodes alliant mise en évidence du polymorphisme et génotypage.....	28
II.V.1. Introduction	28
II.V.2. L'AFLP	29
II.V.3. Les marqueurs RAD	32
II.V.4. L'Ecotilling	35
III. La cartographie génétique	37
III.I. Introduction	37
III.II. Types de population.....	38
III.III. Principe de la cartographie génétique	39

III.III.1.	Assignation des marqueurs à un groupe de liaison.....	39
III.III.2.	Etablissement de l'ordre des marqueurs dans chaque LG	41
III.III.3.	Détermination des distances génétiques.....	41
III.IV.	La cartographie chez la chicorée et chez la laitue.....	42
III.V.	Utilisation des cartes génétiques	43
III.V.1.	Localisation et caractérisation de QTL	43
III.V.2.	La sélection assistée par marqueur.....	44
III.V.3.	L'assemblage des génomes.....	45
IV.	La cartographie comparée.....	47
IV.I.	Introduction	47
IV.II.	Démarche.....	48
IV.II.1.	Choix du niveau taxonomique étudié	48
IV.II.2.	Recherche et utilisation de séquences orthologues	49
IV.III.	Intérêt de la cartographie comparée.....	53
IV.III.1.	Etude de l'évolution des génomes	53
IV.III.2.	Transfert d'informations génétiques	57

I.	But du chapitre	61
II.	Polymorphisme au sein des banques EST	62
II.I.	Les banques EST chez la chicorée	62
II.II.	Mise en évidence de marqueurs microsatellites	63
II.II.1.	Introduction	63
II.II.2.	Matériel et méthodes	63
II.II.2.1.	Création d'une banque d'unigènes.....	63
II.II.2.2.	Recherche de motifs microsatellites	64
II.II.2.3.	Annotation des EST.....	64
II.II.2.4.	Conception des amorces flanquant les SSR	64
II.II.3.	Résultats et discussion	64
II.II.3.1.	Création d'un jeu d'unigènes.....	64
II.II.3.2.	Recherche de motifs microsatellites	66
II.II.3.3.	Sélection des séquences suivant leur fonction.....	71
II.II.3.4.	Sélection des séquences suivant les caractéristiques intrinsèques des microsatellites	74
II.III.	Mise en évidence de SNP	76
II.III.1.	Introduction	76
II.III.2.	Matériel et méthodes	76
II.III.2.1.	Banque de données	76
II.III.2.2.	Analyses bioinformatiques	76
II.III.3.	Résultats et discussion	77
II.III.3.1.	Choix de la banque EST.....	77
II.III.3.2.	Formation et caractérisation des contigs	78
II.III.3.3.	Analyse des SNP.....	78
II.IV.	Conclusion sur l'analyse des banques EST	83
III.	Polymorphisme au sein de séquences orthologues à la laitue	84
III.I.	Introduction	84
III.II.	Matériel et méthodes	85
III.II.1.	Recherche de séquences orthologues entre la laitue et la chicorée.....	85
III.II.1.1.	Banques de données.....	85
III.II.1.2.	Reciprocal best BLAST Hit (RBH)	85

III.II.2.	Mise en évidence de SNP dans les séquences orthologues au sein de la population SAMCHIC.....	85
III.II.2.1.	Population de cartographie SAMCHIC	85
III.II.2.2.	Conception d'amorces et amplification.....	86
III.II.2.3.	Evaluation du polymorphisme par Ecotilling	86
III.II.2.4.	Clonage des produits PCR	87
III.II.2.5.	PCR sur colonie et séquençage	89
III.III.	Résultats et discussion	90
III.III.1.	Recherche de séquences orthologues entre la laitue et la chicorée	90
III.III.2.	Mise en évidence du polymorphisme SNP contenu dans les séquences orthologues au sein du croisement SAMCHIC.....	93
IV.	Polymorphisme au sein de séquences supplémentaires	97
IV.I.	Choix des séquences	97
IV.II.	Analyse du polymorphisme des séquences	99
V.	Conclusion du chapitre.....	99

CHAPITRE 3. GÉNOTYPAGE	101
I. But du chapitre	101
II. Génotypage à l'aide des marqueurs AFLP	102
II.I. Introduction	102
II.II. Matériel et méthodes	103
II.II.1. Ligation des deux brins des adaptateurs	103
II.II.2. Restriction et ligation des adaptateurs	103
II.II.3. Amplification pré-sélective	104
II.II.4. Amplification sélective	105
II.II.5. Migration des fragments AFLP au séquenceur	106
II.II.6. Analyse des sorties AFLP	106
II.III. Résultats et discussion	106
II.III.1. Choix des amorces pour l'amplification sélective	106
II.III.2. Analyse des sorties AFLP	107
II.III.3. Comparaison avec les données obtenues par l'ECAV (UCL)	110
II.IV. Conclusion	110
III. Génotypage à l'aide des marqueurs microsatellites	111
III.I. Introduction	111
III.II. Matériel et méthodes	111
III.II.1. Amplification des motifs SSR à l'aide du système à trois amorces	111
III.III. Résultats et discussion	112
III.III.1. Test d'amplification des amorces microsatellites	112
III.III.2. Génotypage des 200 individus du croisement SAMCHIC	115
III.IV. Conclusion	117
IV. Génotypage à l'aide des marqueurs SNP	118
IV.I. Introduction	118
IV.II. Matériel et méthodes	119
IV.II.1. Méthode d'extension-ligation	119
IV.II.2. Méthode de ligation	120
IV.II.3. Méthode PAMSA	120
IV.II.4. Méthode Bi-PASA	121
IV.III. Résultats et discussion	122
IV.III.1. Mise au point des méthodes de génotypage	122

IV.III.1.1.	Méthode d'extension-ligation	122
IV.III.1.2.	Méthode de ligation	128
IV.III.1.3.	Méthode PAMSA	131
IV.III.1.4.	Méthode Bi-pasa	133
IV.III.1.5.	Récapitulatif des méthodes de génotypage	133
IV.III.2.	Analyse des SNP présents au sein des séquences de chicorée	135
IV.III.2.1.	Analyse des SNP présents au sein des séquences EST de la banque Phytomol	135
IV.III.2.2.	Analyse des SNP présents au sein de séquences orthologues à la laitue et de séquences d'intérêt (carte Cadalen, pathogénèse, floraison, ...)	137
IV.III.2.3.	Analyse du polymorphisme au sein de séquences impliquées dans la synthèse de l'inuline	138
IV.III.2.4.	Récapitulatif des SNP analysés pour le génotypage	139
V.	Conclusion	140

CHAPITRE 4. CONSTRUCTION D'UNE CARTE GÉNÉTIQUE DE LA CHICORÉE	143
I. But du chapitre	143
II. Cartes génétiques du genre Cichorium publiées.	144
III. Matériel et méthodes.....	147
III.I. Jeu de données	147
III.II. Construction de la carte génétique.....	147
III.III. Estimation de la taille du génome.....	148
IV. Résultats et discussion	148
IV.I. Encodage des données génotypiques en vue de la cartographie.....	148
IV.II. Construction de la carte génétique.....	150
IV.III. Caractéristiques de la carte génétique	154
IV.IV. Influence de la taille de la population et du nombre de marqueurs	155
IV.V. Calcul de la taille estimée du génome.....	159
IV.VI. Correspondance cM-paires de base	160
IV.VII. Fonctions associées aux séquences cartographiées.....	160
V. Conclusion	163

CHAPITRE 5. CARTOGRAPHIE COMPARÉE ENTRE <i>LACTUCA</i> ET <i>C. INTYBUS</i> VAR. <i>SATIVUM</i>	165
I. But de ce chapitre	165
II. Introduction	166
III. Matériel et méthodes	169
III.I. Banques de données	169
III.II. Recherche de séquences orthologues entre la laitue et la chicorée	169
III.III. Représentation des paires d'orthologues sur une carte comparative	169
IV. Résultats et discussion	170
IV.I. La cartographie chez la laitue	170
IV.I.1. Cartes génétiques	170
IV.I.2. Carte physique	170
IV.II. Recherche de paires d'orthologues	173
IV.II.1. Stratégie et résultats	173
IV.II.2. Fiabilité des résultats	173
IV.II.2.1. e-Value seuil	173
IV.II.2.2. Couverture du génome total par les banques utilisées	175
IV.III. Analyse de la synténie	179
IV.III.1. Construction de la carte comparative	179
IV.III.2. Analyse des blocs de synténie	179
IV.III.3. Hypothèse sur le degré de synténie	183
IV.III.4. Transfert de données entre la chicorée et la laitue	186
V. Conclusion	191
CONCLUSION ET PERSPECTIVES	193
BIBLIOGRAPHIE	199

Liste des abréviations

A	Adénine
ADN	Acide DésoxyriboNucléique
ADNc	Acide DésoxyriboNucléique complémentaire
ADNg	Acide DésoxyriboNucléique génomique
AFLP	Amplified Fragments Length Polymorphism
ARNm	Acide RiboNucléique messenger
ASO	Allele-Specific Oligonucleotide
BAC	Bacterial Artificial Chromosome
Bi-PASA	Bidirectional PCR amplification of Specific Alleles
BLAST	Basic Local Alignment Search Tool
C	Cytosine
Cds	Codant
CGP	Compositea Genome Project
CGPDB	Compositea Genome Project DataBase
cM	centiMorgan
ECAV	Unité d'Ecophysiologie et d'Amélioration végétale
EST	Expressed Sequence Tag
FEH	Fructan Exo-Hydrolase
FFT	Fructan Fructan fructosylTransferase
FUNDP	Facultés Universitaires Notre Dame de la Paix
G	Guanine
Gb	Gigabase
GBS	Genotyping-By-Sequencing
GO	Genes Ontology
kb	kilobase
LG	Groupe de Liaison
LSO	Locus-Specific Oligonucleotide
Mb	Million de bases
MYA	Million Years Ago
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
ORF	Open Reading Frame
PAMSA	PCR Amplification of Multiple Specific Alleles

pb	paire de base
PCR	Polymerase Chain Reaction
QTL	Quantitatif Trait Loci
RAD	Restriction-site Associated DNA
RAPD	Random Amplification of Polymorphic DNA
RBH	Reciprocal best BLAST Hit
RFLP	Restriction Fragment Length Polymorphism
RIL	Recombinant Inbred Line
SAMPL	Selective Amplification of Microsatellite Polymorphic Loci
SCFA	Short-Chain Fatty Acids
SNP	Single Nucleotide Polymorphism
SSCP	Single Strand Conformation Polymorphism
SSR	Simple Sequence Repeats
SST	Sucrose Sucrose fructosylTransferase
STS	Sequence Tag Site
T	Thymine
Tm	melting Temperature
UCL	Université Catholique de Louvain
URBV	Unité de Recherche en Biologie Végétale
UTR	UnTranslated Region
WGD	Whole Genome Duplication

Introduction générale

La chicorée industrielle, *Cichorium intybus* var. *sativum*, est une espèce de la famille des Asteraceae cultivée dans le but de produire de l'inuline, un polysaccharide aux vertus prébiotiques, utilisé dans l'industrie agro-alimentaire.

Afin d'augmenter la rentabilité d'espèces cultivées, les sélectionneurs créent continuellement de nouvelles variétés de plus en plus performantes via une sélection principalement phénotypique. Depuis une vingtaine d'années, les outils moléculaires sont venus renforcer les méthodes traditionnelles. Ces outils sont utilisés pour la recherche de liens de parenté, l'analyse de la diversité génétique et des distances génétiques, ainsi que pour la sélection assistée par marqueurs.

Afin de créer de nouveaux outils moléculaires pour la sélection de la chicorée, nous avons analysé les banques EST disponibles pour cette espèce dans le but de développer de nouveaux marqueurs co-dominants et spécifiques de locus de types microsatellite et SNP.

Ces marqueurs, associés à des séquences exprimées, ont été utilisés pour la construction d'une carte génétique en collaboration avec l'UCL et la société Cosucra. Une carte génétique permet d'avoir une meilleure connaissance de l'organisation du génome de la chicorée et s'avère donc être un outil intéressant dans différentes applications telles que l'analyse de QTL, l'introgession ou encore la création de carte physique.

Afin d'augmenter les sources de données génétiques disponibles pour la chicorée, une cartographie comparée avec la laitue, une espèce phylogénétiquement proche, a été réalisée. Le but de cette étape était d'analyser la synténie c'est-à-dire la conservation de l'organisation des génomes entre les deux espèces afin de voir s'il était possible de transférer chez la chicorée, les données génétiques disponibles pour la laitue telle que la localisation sur le génome de séquences ou de QTL (Quantitative traits loci).

Le premier chapitre de ce manuscrit est composé d'une revue bibliographique afin de replacer ce travail dans son contexte.

Le chapitre 2 décrit les différentes démarches ainsi que les résultats obtenus pour la mise en évidence de microsatellites et SNP au sein de séquences EST.

Dans le chapitre 3, les différentes méthodes de génotypage mises au point ainsi que l'utilisation des marqueurs SNP et microsatellites pour le génotypage de la population de ségrégation sont détaillées.

La construction de la carte génétique de la chicorée réalisée en collaboration avec l'ECAV de l'UCL et la société Cosucra est expliquée dans la chapitre 4.

Enfin, le dernier chapitre est consacré à la cartographie comparée entre la chicorée et la laitue et à son utilisation pour le transfert de données génétiques localisées sur le génome de la laitue telles que des QTL et des loci phénotypiques.

Chapitre 1

Revue bibliographique

Chapitre 1. REVUE BIBLIOGRAPHIQUE

I. CICHORIUM INTYBUS VAR. SATIVUM : CLASSIFICATION, INTÉRÊT ET SÉLECTION

Lors de cette thèse, un intérêt particulier a été porté au développement d'outils moléculaires avantageux pour l'amélioration des variétés de chicorée industrielle. Dans ce premier chapitre, cette espèce est replacée dans son contexte taxonomique et agronomique dans le but de cerner les problématiques liées à la culture de la chicorée.

I.I. CLASSIFICATION TAXONOMIQUE DE LA CHICORÉE INDUSTRIELLE

Règne : Plantae			Famille: Asteraceae (ou Compositae)
Sous-règne : Tracheobionta			Sous-famille: Cichorioideae
Division : Magnoliophyta			Tribu: Cichorieae
Classe : Magnoliopsida			Genre : Cichorium
Sous-classe : Asteridae			Espèce : intybus
Ordre : Asterales			Variété : sativum

La chicorée industrielle fait partie de la grande famille des Asteraceae (Compositae). Cette famille comporte entre 24.000 et 30.000 espèces soit près de 15 % des plantes à fleurs.

La sous-famille des *Cichorioideae* est composée de cinq tribus donc la majorité trouve son origine au sein du continent africain (Funk et al. 2005). La tribu *Cichorieae* (ou *Lactuceae*), originaire de l'Afrique du Nord, de la Méditerranée et du sud de l'Europe, comporte plusieurs

genres tels que *Lactuca* dont fait partie la laitue, *Taraxacum* regroupant les différentes espèces de pissenlit et *Cichorium* (**figure 1.1**) (Vermeulen et al. 1994).

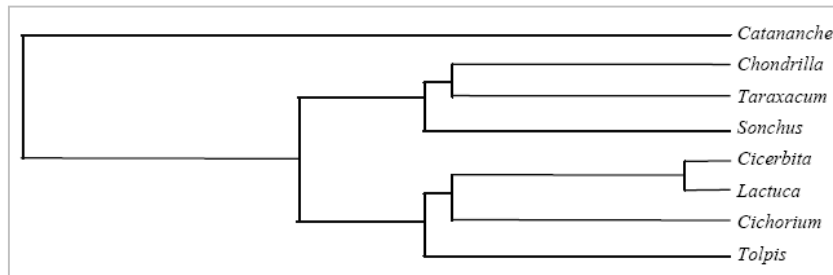


Figure 1.1 : Dendrogramme UPGMA des différents genres de *Chicorideae* (*Lactuceae*) basé sur le polymorphisme de l'ADN mitochondrial. (Vermeulen et al. 1994)

Le genre *Cichorium* comprend deux espèces cultivées : *C. intybus* (la chicorée) et *C. endivia* (l'endive) et quatre espèces sauvages : *C. bottae*, *C. spinosum*, *C. calvum* et *C. pumilum*. *C. intybus* englobe le groupe « sylvestre » (la chicorée sauvage), le groupe « foliosum » (la chicorée witloof ou chicon), le groupe « pain de sucre » que l'on retrouve en Italie, le groupe « radicchio » et enfin, le groupe « sativum » c'est-à-dire la chicorée industrielle ou à café (Kiers et al. 2000).

En 2011, 10 variétés cultivées de chicorée industrielle étaient inscrites au catalogue belge: Canzona, Continuo, Crescendo, Diesis, Enigme, Hera, Maurane, Melci, Orchies et Zingaro (Pannecouque et al. 2011)

I.II. INTÉRÊT DE LA CHICORÉE INDUSTRIELLE

I.II.1. Historique de la culture de chicorée et production

Les premières traces de chicorée remontent à l'Égypte ancienne où celle-ci était utilisée comme plante digestive et dépurative. Les racines étaient découpées et mangées en apéritif. Au VIII^{ème} siècle, Charlemagne rendit la culture de chicorée obligatoire dans les jardins médicaux des châteaux. Plus tard, lors des guerres de la révolution française, les racines de chicorée furent torréfiées afin d'obtenir une décoction semblable à du café dont le prix était en continuelle croissance. La culture de chicorée dite « à café » s'est ensuite développée dans le nord de la France, en Belgique ainsi qu'en Autriche et en Allemagne (Doré and Varoquaux 2006). A ce jour, la

chicorée est toujours cultivée pour la torréfaction de ses racines mais également pour la production d'inuline.

La culture de chicorée industrielle pour l'extraction d'inuline est relativement récente (elle date des années '80). Les trois pays qui cultivent cette espèce afin de produire de l'inuline sont la Belgique, les Pays-Bas et le Chili, dont la superficie pour cette culture était en 2010 respectivement de 8.027, 4.691 et 3.186 ha (www.statbel.fgov.be, www.cbs.nl, www.ine.cl).

La Belgique est donc le premier producteur de chicorée industrielle avec plus de 50% de la production mondiale (en ha). De 1998 à 2010, la surface cultivée en Belgique a diminué de moitié. Cependant, les modes de sélection ainsi que l'amélioration des variétés ont permis d'augmenter le rendement de 22% sur la même période (**figure 1.2**).

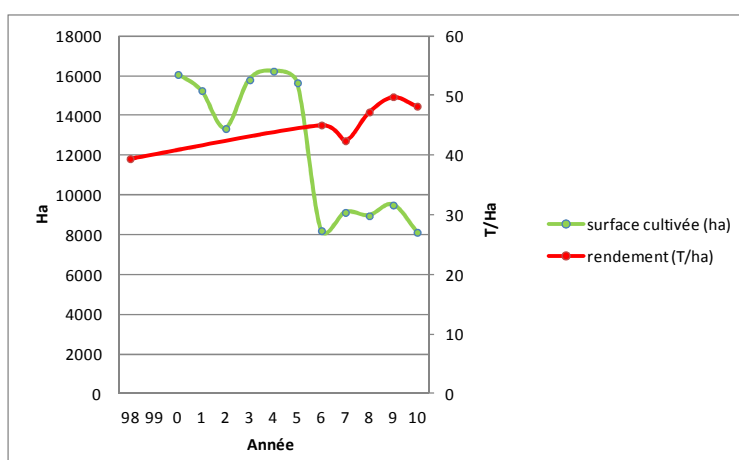


Figure 1.2 : Surface cultivée (en ha) et rendement (en T/ha) de la chicorée industrielle en Belgique de 1998 à 2010 (Source : SPF Economie)

I.II.2. L'inuline : définition, synthèse, rôle et utilisation

▪ Fructane et inuline : définition

La chicorée industrielle tout comme 15 % des plantes à fleurs utilise, en plus de l'amidon, les fructanes comme réserve de carbohydrates au niveau de la vacuole. Un fructane est une oligo- ou polysaccharide contenant au moins deux unités de fructose adjacent. Cinq structures différentes de fructanes ont été observées chez les plantes

supérieures : le type inuline (chaîne de fructose en liaison β 2-1 de type G1-2F1-2Fn), le type lévane (chaîne de fructose en liaison β 2-6 de type G1-2F6-2Fn), le type néosérie inuline (chaîne de fructose en liaison β 2-1 où le glucose se trouve entre deux fructoses) , le type néosérie lévane (chaîne de fructose en liaison β 2-6 où le glucose se trouve entre deux fructoses) et le type lévane mélangé (contient des liaisons de type β 2-1 et β 2-6 entre les fructoses) (Vijn and Smeekens 1999). Les fructanes de type inuline sont principalement présents dans les espèces appartenant à la famille des Astéraceae telles que le topinambour (*Helianthus tuberosus*), l'artichaut (*Cynara scolymus*), le pissenlit (*Taraxacum officinale*) et bien évidemment la chicorée (Van Laere and Van Den Ende 2002).

- Métabolisme de l'inuline (**figure 1.3**)

La synthèse de l'inuline se fait via deux enzymes principales : 1-SST et 1-FFT. La 1-SST (Sucrose-1-Sucrose Fructosyltransferase) transfère un fructose provenant d'un saccharose sur le carbone 1 d'un autre saccharose. La 1-FFT (Fructose-1-Fructose Fructosyltransferase) transfère le fructose d'un fructane sur un autre fructane. La 1-SST permet donc l'initiation de la synthèse du 1-Kestose et la 1-FFT permet d'en augmenter le degré de polymérisation. La dépolymérisation de l'inuline est réalisée par des enzymes appelées FEH (Fructan ExoHydrolase I, IIa et IIb). Celles-ci vont détacher les fructoses de la chaîne d'inuline par hydrolyse. (Edelman and Jefford 1968; Van Laere and Van Den Ende 2002)

Lors de la phase de croissance de la chicorée (de juillet à septembre), l'activité conjointe de la 1-SST et de la 1-FFT augmente. Cela implique une accumulation d'inuline possédant un degré de polymérisation (DP) élevé au sein de la vacuole. Lorsque la température commence à diminuer et que la phase de croissance est terminée, on observe une diminution de l'activité de la 1-SST et une augmentation de l'activité des FEH. La 1-FFT reste constante. Lors de cette période, les FEH hydrolysent les fructanes, ce qui implique une diminution de leur degré de polymérisation et une augmentation de la concentration en fructose libre. En parallèle, les 1-FFT, en absence de 1-SST, ont la capacité d'ajouter sur le saccharose et sur le fructose libre des fructoses provenant de fructanes. Cela implique une diminution du DP moyen ainsi qu'une augmentation de la concentration en fructanes dans la vacuole (Van Laere and Van Den Ende 2002).

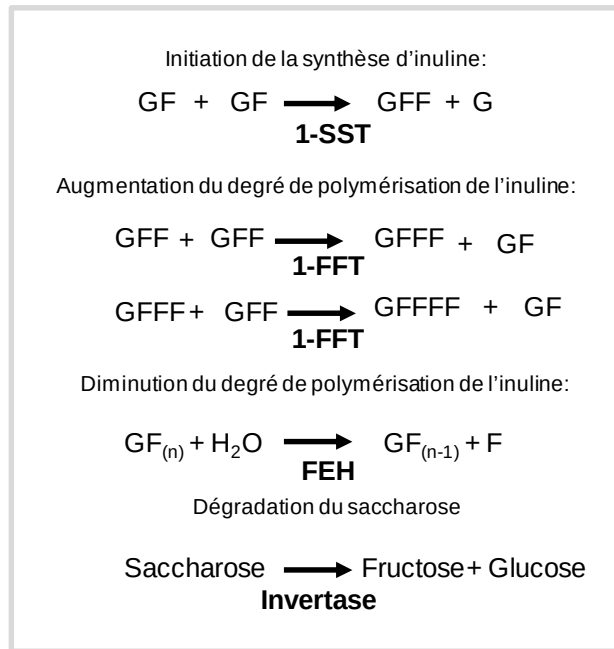


Figure 1.3 : Enzymes intervenant dans la synthèse et la dégradation de l'inuline (fructan de forme G(F)_n). G : glucose. F : fructose. La 1-SST (sucrose-1-sucrose fructosyle transférase) catalyse le transfert d'un fructose provenant d'un saccharose (GF) sur le carbone 1 d'un autre saccharose formant un 1-kestose (GFF). Cette enzyme permet donc l'initiation de la synthèse de l'inuline. La 1-FFT (fructose-1-fructose fructosyltransférase) catalyse le transfert d'un fructose provenant d'un fructane de forme G(F)_n sur une autre fructane. Cette enzyme permet donc d'augmenter le degré de polymérisation de l'inuline. Les FEH (fructane exohydrolase) hydrolysent un fructose de l'inuline. Cette enzyme diminue donc le degré de polymérisation. De manière plus générale, les invertases hydrolysent le saccharose en glucose et en fructose

▪ Rôle physiologique des fructanes chez les plantes

Il existe divers avantages évolutifs qui peuvent expliquer la présence des fructanes chez les plantes. Tout d'abord, les fructanes sont des polysaccharides de réserve solubles dans l'eau dont les ressources peuvent être rapidement mobilisées grâce à l'action de différentes enzymes. Ils semblent donc adaptés aux changements brusques. De plus, les enzymes impliquées dans leur biosynthèse sont moins sensibles au froid que les enzymes impliquées dans le métabolisme de l'amidon. En-dessous d'une température de 10°C , la synthèse des fructanes est toujours active. Cela empêche une accumulation de saccharose dans la cellule et donc une rétroinhibition de la photosynthèse (Pollock 1986).

Un des avantages majeurs des fructanes est lié à la dépolymérisation et à l'augmentation de la concentration des fructanes en fin de croissance de la plante lorsque les températures commencent à diminuer. Ce mécanisme apporte une protection contre le gel et la dessiccation de la plante : une augmentation de la concentration en fructanes augmente d'abord la pression osmotique au sein de la cellule et empêche la plasmolyse (Livingston et al. 2009). Ensuite, il semblerait qu'une partie des fructanes soit déplacée au niveau de l'apoplasme. La concentration élevée de ces sucres autour de la cellule permettrait de diminuer le point de congélation de l'eau pariétale (Livingston and Henson 1998). Enfin, les fructanes jouent également un rôle dans la stabilisation des membranes afin d'éviter une fuite du cytoplasme due à l'augmentation de la perméabilité des cellules en cas de déshydratation (Van Laere and Van Den Ende 2002).

- Utilisation de l'inuline dans l'industrie agro-alimentaire

Afin d'être utilisée dans l'agro-alimentaire, l'inuline est extraite de la racine de chicorée, comme le saccharose pour la betterave, par un système de diffusion due à une dénaturation des membranes induite par de l'eau chaude à plus de 80°C (Arzate 2005). L'inuline ainsi extraite peut être séchée et utilisée comme telle. Une étape supplémentaire d'hydrolyse partielle peut être réalisée à l'aide d'endoinulase pour former des oligofructoses de degrés de polymérisation plus faibles (Rocha et al. 2006).

L'inuline extraite et séchée se présente sous l'aspect d'une poudre blanche au goût neutre possédant peu ou pas de pouvoir sucrant (10 % par rapport au saccharose) et peu soluble dans l'eau (maximum 10 % à température ambiante). L'inuline, une fois mélangée à l'eau, forme une gelée blanche à la structure crémeuse qui peut être facilement ajoutée à divers nutriments dans le but de remplacer les graisses. Les oligofructoses sont beaucoup plus solubles (80% à température ambiante) et possèdent un pouvoir sucrant plus prononcé (35 % par rapport au saccharose). Généralement associés à des édulcorants, ces sucres sont utilisés comme substituts du saccharose (Franck 2002).

Après hydrolyse totale de l'inuline, on obtient un sirop de fructose utilisé comme édulcorant avec un pouvoir sucrant de 1,1 à 1,5 fois supérieur au saccharose (Lecerf 2009). Depuis 2006, les quotas accordés par l'Organisation Commune du Marché (OCM) à la vente de

fructose ont fortement diminué (-40 %) (ICEDD 2007). La marge bénéficiaire obtenue pour la vente de fructose n'est plus rentable. A ce jour, ce produit issu de l'inuline n'est donc plus commercialisé.

- Caractéristiques prébiotiques de l'inuline

L'inuline ainsi que les oligofructanes jouent un rôle prébiotique dans le système digestif. Les prébiotiques sont des aliments non digestibles par l'homme qui stimulent la croissance ou l'activité de bactéries présentes dans la flore intestinale et bénéfiques pour la santé. L'inuline et les oligofructanes sont des substrats qui activent spécifiquement la fermentation des *Bifidobacteriae* et des *Lacobacilli*, deux types de bactéries présentes dans le colon et ayant un effet bénéfique sur la santé (Gibson and Roberfroid 1995).

Plusieurs mécanismes peuvent expliquer l'impact de ces bactéries sur la santé. Tout d'abord, l'activation de la fermentation par les prébiotiques implique la production d'acides gras à courte chaîne (SCFA, Short-Chain Fatty Acids). Ces molécules sont absorbées par l'organisme hôte. L'inactivation de la glycogénèse hépatique par ces substances va avoir une influence sur le taux de sucre dans le sang. De plus, Il semblerait que le taux de cholestérol et de triglycéride soit également influencé par la présence de SCFA (Kaur and Gupta 2002; Ooi et al. 2010; Roberfroid et al. 2010) (Kaur and Gupta 2002). D'autres produits de la fermentation semblent inhiber l'activité d'enzymes carcinogènes et donc pourraient jouer un rôle dans la prévention du cancer du côlon (Pool-Zobel and Sauer 2007).

La présence de prébiotiques au sein du côlon modifie également la nature de la flore intestinale. En effet, la prolifération de *Bifidobacteriae* engendre la production par celles-ci de bactériocines ainsi qu'une diminution du pH du milieu. Ces deux facteurs ont un impact négatif sur la prolifération de certaines bactéries pathogènes telles que *Escherichia coli* fimbria de type 1 ou salmonellae, ce qui permet de diminuer les troubles gastro-intestinaux. La modification de la flore intestinale semble également stimuler le système immunitaire de l'organisme hôte, permettant donc une meilleure défense de l'organisme (Gibson et al. 2005).

I.III. LA SÉLECTION CHEZ LA CHICORÉE INDUSTRIELLE

Depuis les années '80, les sélectionneurs ont développé des dizaines de variétés de chicorée industrielle dans le but de trouver un compromis entre différents critères de sélection (Doré and Varoquaux 2006). On peut classer ces critères en quatre catégories : le rendement en sucre, la résistance aux maladies et aux herbicides, la morphologie racinaire et la résistance à la montaison.

Le rendement en sucre est évidemment important. Afin d'augmenter la concentration des fructanes à haut et moyen degré de polymérisation, les variétés peu sensibles à la dépolymérisation sont privilégiées.

La chicorée industrielle fait face à plusieurs maladies foliaires telles que l'oïdium (*Erysiphe cichoracearum*), qui se manifeste par l'apparition d'un duvet blanchâtre recouvrant les feuilles de chicorée, ou encore la rouille (*Puccinia cichorii*), qui se caractérise par l'apparition de spores sous forme de pustules de couleur brun-noir sur les feuilles. D'autres maladies foliaires telles que la sclérotiniose (*Sclerotinia sclerotiorum*), également appelée « pourriture blanche », ou encore l'alternariose (*Alternaria sp*) sont également rencontrées chez la chicorée industrielle (Christine Notte, Chicoline, Cosucra groupe Warcoing S.A, communication personnelle). Plusieurs genres de champignons pathogènes racinaires ont également été identifiés : *Fusarium*, *Phoma*, *Cylindrocarpon* et *Mucor* (Vincent Dielen, Chicoline, Cosucra groupe Warcoing S.A, communication personnelle). Afin d'éviter au maximum une infection des feuilles et des racines de chicorée, les variétés sont sélectionnées en tenant compte de leur résistance aux différentes maladies.

La sélection de plantes tolérantes à un herbicide est également un avantage pour la culture, le traitement pouvant ainsi être précoce sans endommager la culture (Manderyck 2011).

La morphologie racinaire est un critère important en sélection : des racines fourchues et filiformes ont plus de chance d'être cassées lors de l'arrachage que des racines trapues et bien droites. De plus, la tare terre c'est-à-dire l'ensemble des matières étrangères récupérées lors de la récolte est aussi dépendante de la forme de la racine. Le sélectionneur privilégiera donc des racines denses et bien droites possédant peu de fourches.

La chicorée industrielle est une plante bisannuelle nécessitant une vernalisation et donc une saison froide pour monter et fleurir. Cependant, pour les besoins de l'industrie, cette plante est cultivée comme une plante annuelle afin d'éviter la montaison. En effet, la production de graines dans les champs engendre une « pollution » pour la culture suivante. De plus, la montaison génère une diminution des réserves de sucre dans la racine puisque celles-ci sont utilisées pour le développement de l'inflorescence. Enfin, lors de la montaison, les racines deviennent fibreuses, ce qui rend leur traitement plus difficile (Dielen et al. 2005). Pour toutes ces raisons, les sélectionneurs tentent de développer des variétés peu propices à la montaison.

Les méthodes classiques utilisées pour discerner et retenir les individus d'intérêt pour les différents critères décrits ci-dessus reposent sur des observations phénotypiques longues et laborieuses nécessitant, suivant le caractère analysé, une plante à un stade de croissance avancé (Doré and Varoquaux 2006). Depuis une vingtaine d'années, le développement de marqueurs moléculaires a permis l'accélération des méthodes classiques d'évaluation des différents cultivars.

II. LES MARQUEURS MOLÉCULAIRES

II.1. INTRODUCTION

Avant toute chose, il est nécessaire de définir certains termes :

- Un marqueur moléculaire est un marqueur génétique également appelé locus marqueur. Un locus marqueur est un locus polymorphe qui renseigne sur le génotype de l'individu qui le porte (de Vienne 1998).
- Un marqueur co-dominant est un marqueur qui permet l'identification des deux allèles d'un individu diploïde (allèle A et allèle B). L'hétérozygote AB peut donc être distingué des homozygotes AA et BB. Un marqueur dominant ne met en lumière qu'un seul des deux allèles (par exemple l'allèle A) de l'individu génotypé. Il n'est donc pas possible de discerner l'hétérozygote AB de l'homozygote AA.

- Un marqueur spécifique de locus ou non-anonyme renseigne sur le génotype d'un locus dont la séquence est connue. Un marqueur anonyme renseigne sur le génotype d'un locus dont la séquence n'est pas connue.
- Les marqueurs moléculaires peuvent être également classés suivant leur localisation. Un marqueur non-génique sera localisé au sein des régions non transcrites du génome, tandis qu'un marqueur génique se trouvera au sein d'un gène.

Le développement de marqueurs moléculaires a commencé au début des années '80 avec l'arrivée de la RFLP (Restriction Fragment Length Polymorphism) (Botstein et al. 1980). Cette méthode repose sur la mise en évidence du polymorphisme de longueur de fragments obtenus suite à l'action d'enzymes de restriction couplée à l'hybridation d'une sonde nucléotidique.

L'arrivée de la technique PCR a permis de développer de nouvelles techniques de génotypage. Les premières méthodes permettaient principalement d'utiliser des marqueurs anonymes grâce aux techniques MAAP (Multiple Arbitrary Amplicon Profiling). Ce terme générique désigne toutes les techniques d'amplification PCR qui utilisent des amorces arbitraires pour générer des profils d'amplification. Parmi celles-ci, on retrouve l'AFLP (Amplified Fragment Length Polymorphism), une méthode utilisée fréquemment chez de nombreuses espèces végétales. Enfin, l'arrivée des techniques de séquençage a permis de développer de nouveaux types de marqueurs moléculaires tels que les marqueurs microsatellites et les marqueurs SNP (Single Nucleotide Polymorphism). A ce jour, le séquençage à haut débit permet de mettre en évidence et de génotyper un très grand nombre de marqueurs répartis sur la totalité du génome.

Dans un premier temps, le polymorphisme à l'origine des marqueurs microsatellites et SNP sera décrit. Ensuite, nous détaillerons comment mettre en évidence ce polymorphisme et comment l'utiliser pour le génotypage. Enfin, trois méthodes permettant de réaliser de façon simultanée la mise en évidence du polymorphisme et le génotypage seront expliquées.

II.II. LES MICROSATELLITES

II.II.1. Introduction

Il existe plusieurs types de séquences répétées au sein du génome. Celles-ci peuvent être classées en différentes catégories suivant leur taille (Tautz 1993)

- **Satellites** : segments répétés de 100 nucléotides ou plus formant un fragment plus ou moins uniforme de 10^3 - 10^7 nucléotides de long.
- **Minisatellites** : segments répétés de 10-100 nucléotides formant un fragment plus ou moins uniforme de 10^2 - 10^5 nucléotides de long.
- **Microsatellites** : segments répétés de 2 à 6 nucléotides formant un fragment plus ou moins uniforme d'environ 10^2 nucléotides de long.
- **Fragments mononucléotidiques** : série uniforme de mononucléotides de n'importe quelle taille.

Lors de cette thèse, un intérêt particulier a été porté aux microsatellites. En effet, ces séquences répétées, de par leur polymorphisme élevé, sont des marqueurs moléculaires intéressants.

Différentes abréviations désignent ces régions répétées telles que les STR (Short tandem repeats), SSR (Simple Sequence Repeats), plus rarement, VNTR (Variable Number of Tandem Repeats) ou encore SSLP pour Simple Sequence Length Polymorphism.

Nous allons tenter de comprendre comment et pourquoi ces séquences sont présentes au sein du génome et plus particulièrement dans les régions transcrites puisque c'est dans celles-ci que nous les identifions.

II.II.2. Mécanismes d'évolution et taux de mutation des microsatellites

Les mécanismes responsables de l'évolution des microsatellites sont détaillés ci-dessous afin de mieux comprendre la provenance du polymorphisme de ces séquences.

L'hypothèse la plus souvent défendue pour expliquer l'évolution du nombre de répétitions dans un microsatellite est appelée « slippage replication » (**Figure 1.4**) (Levinson and Gutman 1987). Après réplication, l'extrémité 3' du brin nouvellement synthétisé se réhybride avec une région complémentaire en amont du brin matrice. Il y a ainsi formation d'une boucle dans le nouveau brin. La polymérase va donc ajouter une unité de répétition supplémentaire. Il se peut également que la boucle se forme au niveau du brin matrice. La polymérase synthétisera donc une répétition en moins.

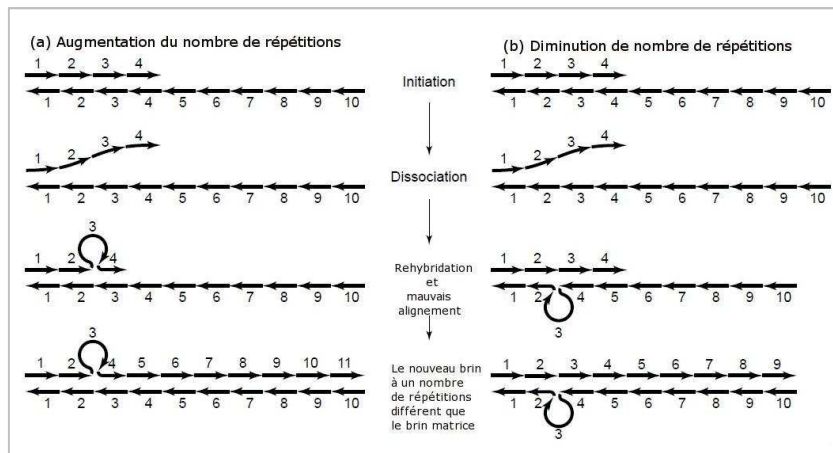


Figure 1.4 : Model de « slippage replication » des microsatellites. Les unités de répétition sont représentées par des flèches. Les chiffres représentent le numéro de l'unité de répétition sur chaque brin d'ADN. (a) Lors de la réplication d'une région répétée, les deux brins d'ADN sont dissociés. Lors de la rehybridation, il peut y avoir un mauvais alignement des séquences. Le brin nouvellement synthétisé peut s'aligner de façon incorrecte avec une unité de répétition en aval du brin matrice. Une boucle se forme sur le nouveau brin qui possède au final une répétition supplémentaire. (b) Même principe que (a) mais l'alignement incorrect forme une boucle au niveau du brin matrice. Le nouveau brin possède donc au final une répétition en moins (Ellegren, 2000)

Une autre hypothèse a également été proposée pour expliquer l'évolution des SSR : la conversion de gènes lors de la recombinaison serait également une source de variation du nombre de répétitions au

sein des microsatellites. La conversion de gènes représente la conversion non réciproque entre deux chromosomes homologues, ce qui induit la perte d'un des deux allèles. Ce modèle est généralement accepté pour les séquences répétées de type minisatellite (Richard et al. 2008). Concernant les microsatellites, deux observations sont en défaveur de cette hypothèse. Tout d'abord, les régions bordantes des microsatellites mutés ne montrent pas de recombinaison. De plus, la distribution et les motifs microsatellites sont identiques sur les chromosomes hemizygotes comme le chromosome Y, par rapport aux autres chromosomes, ce qui suggère que les événements de mutation ne nécessitent pas de contact entre les chromosomes homologues (Ellegren 2000). Cependant, il a été montré de manière expérimentale que la conversion de gène peut modifier le nombre de répétitions au sein de microsatellites de type trinuécléotide (Richard and Paques 2000). Il semblerait donc que le modèle de « slippage replication » soit la source principale de variation du nombre de répétitions au sein des microsatellites mais qu'en parallèle et de façon plus rare, d'autres mécanismes rentrent également en jeu.

L'évolution des microsatellites semble suivre un schéma qui implique une expansion du nombre de répétitions jusqu'à arriver à une taille critique où le microsatellite se contracte. Cette évolution particulière est fonction du rapport insertion/délétion lors du « slippage replication ». De plus, l'apparition de mutations ponctuelles dans les microsatellites de grande taille va interrompre et donc également modifier la séquence répétée (Ellegren 2000; Richard et al. 2008).

Le taux de mutation chez les microsatellites est compris entre 10^{-2} et 10^{-6} par locus et par génération. Celui-ci est très variable d'une espèce à l'autre mais également au sein même d'une espèce et entre loci. La longueur du microsatellite semble influencer directement le taux de mutation. Les séquences répétées de grandes tailles ont un taux de mutation plus élevé. Cela peut s'expliquer par le fait que la probabilité qu'un « slippage » ait lieu est plus grande quand le nombre de répétitions est plus élevé (Ellegren 2000). Enfin, la variabilité du taux de mutation entre les espèces pourrait en partie être expliquée par une différence d'efficacité de la réparation des erreurs lors de la réplication de l'ADN (Li et al. 2004).

II.II.3. Les microsatellites dans les régions transcrites

Les marqueurs microsatellites (ou SSR) développés dans cette thèse sont localisés au sein de régions transcrites. Nous allons décrire les caractéristiques de ces séquences répétées localisées dans cette région précise du génome.

- Fréquences des SSR dans les régions transcrites

Les microsatellites présents au sein des régions transcrites sont mis en évidence grâce à la construction et l'analyse de banques EST (Expressed Sequence Tag). Une banque EST est une banque de séquences provenant de la rétrotranscription d'ARNm. Généralement, les EST sont présentes en plusieurs copies dans la banque. Afin d'éviter les redondances lors de la recherche de motifs microsatellites, les EST sont regroupées en « contigs » et seule la séquence consensus est retenue. Au final, on obtient une banque d'unigènes, c'est-à-dire une banque de séquences uniques.

L'analyse de banques EST chez différentes espèces de plantes montre qu'environ 5 % des EST présentent des SSR avec une taille minimum de 20 pb (Varshney et al. 2005).

Plusieurs recherches de motifs SSR au sein de banques EST d'Asteraceae ont été réalisées. Une banque EST publiée en 2009 montre que 4,5 % des unigènes de *Lactuca sativa* et de *L. serriola* présentent des SSR de minimum 20 pb dont l'unité de répétition est comprise entre 2 et 8 nucléotides (Simko 2009). Cinq autres banques EST provenant également du genre *Lactuca* montrent que suivant l'espèce, 6,23 à 7,87 % des unigènes possèdent des SSR de minimum 12 nucléotides dont l'unité de répétition est comprise entre 1 et 7 nucléotides (Riar et al. 2011). Ces résultats semblent logiques. En effet, la sélection de SSR plus petits lors de la deuxième étude explique le pourcentage plus élevé de SSR détectés.

Chez *Carthamus tinctorius*, une autre Astéracée, 22,8 % des unigènes d'une banque EST contiennent des SSR possédant minimum cinq répétitions pour les motifs dinucléotidiques et quatre répétitions pour les autres motifs. Seuls les SSR de types di-, tri- et tétranucléotidiques ont été retenus (Chapman et al. 2009). Ce pourcentage représente plus du double de ce qui est observé chez le tournesol lors d'une analyse réalisée dans des conditions semblables. En effet, 10,9 % des

17.904 unigènes analysés de *Helianthus annuus* possèdent des motifs SSR dont le nombre de répétitions di-, tri- et tétranucléotidiques est supérieur ou égal à cinq (Heesacker et al. 2008). Les motifs SSR au sein des EST de *Carthamus tinctorius* semblent donc plus nombreux qu'au sein du genre *Lactuca* et chez le tournesol.

Enfin, chez la chicorée industrielle, l'analyse d'une banque EST composée de 5.940 unigènes montre que 15 % de ces séquences possèdent des SSR de types di-, tri, tétra- et pentanucléotidiques de minimum 12 pb (Dauchot 2007). Les motifs SSR semblent donc plus nombreux dans les EST de chicorée analysées qu'au sein des EST du genre *Lactuca*. Les valeurs semblent, au vue des différences d'analyses, proches de celles obtenues pour le tournesol (Heesacker et al. 2008).

- Types de motifs SSR dans les régions transcrites

De manière générale, les microsatellites de type trinuéclotide sont plus fréquemment représentés dans les régions transcrites. En effet, la variation du nombre de répétitions au sein de ces SSR ne modifie pas le cadre de lecture des gènes, ce qui permet aux SSR d'évoluer sans pour autant modifier drastiquement la séquence et donc la structure de la protéine codée par cette séquence (Katti et al. 2001; Li et al. 2004). Chez la chicorée, les SSR de type trinuéclotide sont effectivement les plus abondants (55 %), suivis de loin par les dinuéclotides (19 %), les tétranuéclotides (17 %) et les pentanucléotides (9 %). La **figure 1.5** résume la distribution des différents types de microsatellites au sein de cette banque d'unigènes (Dauchot 2007).

D'une espèce à l'autre, le motif trinuéclotidique le plus fréquent varie fortement. Chez *Arabidopsis*, le motif AAG est le plus abondant (Li et al. 2004) alors que chez la chicorée, le motif ATC est le motif le plus observé (Dauchot 2007). La variation du nombre de répétitions des microsatellites présents au sein même des régions codantes va modifier la teneur en acides aminés de la protéine. Cependant, vu la grande diversité des motifs SSR chez les différentes espèces, il est difficile d'y voir une tendance particulière.

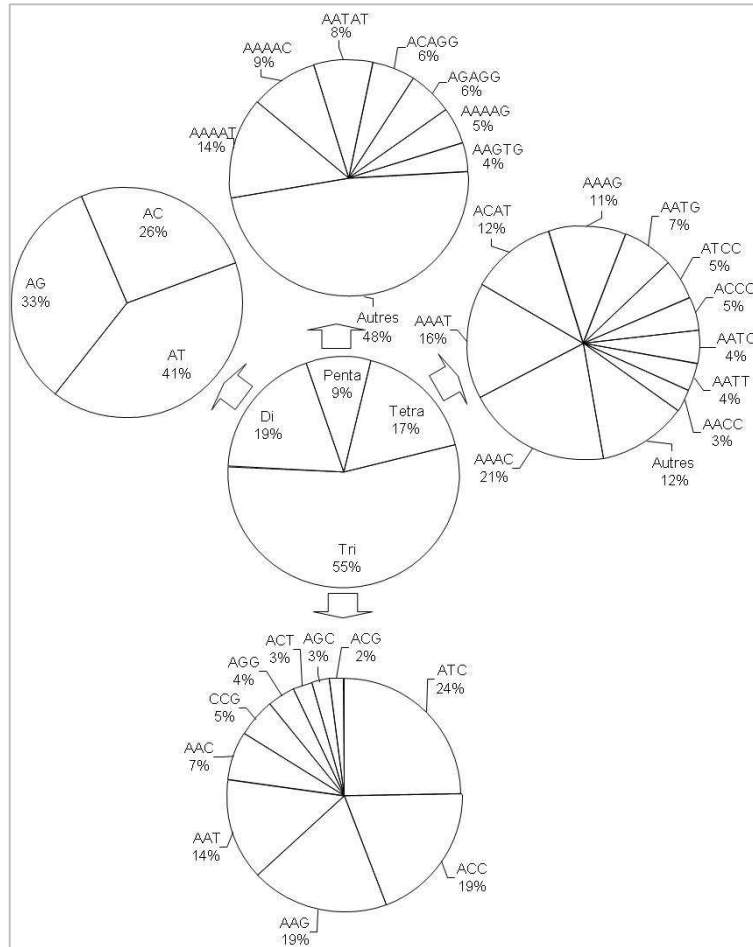


Figure 1.5 : Abondance relative des motifs SSR au sein d'EST provenant de racines de chicorée industrielle. 10.481 EST ont permis de former 5.940 unigènes potentiels. 881 unigènes contiennent 1.076 SSR dont la longueur totale est $\geq$ à 12 pb (Dauchot 2007)

- Distribution des SSR dans les différentes zones des régions transcrites

Les régions transcrites sont composées d'introns et d'exons. Dans les exons, on retrouve les régions 5'- et 3'-UTR non traduites et les régions codantes. Chez *Arabidopsis* ainsi que chez le riz, les SSR sont plus fréquents dans les régions 5'-UTR par rapport aux régions codantes, aux introns et aux régions 3'-UTR. Cette abondance de séquences répétées pourrait s'expliquer par le rôle de régulation de la transcription et de la traduction de ces SSR (Fujimori et al. 2003). Par

contre, il y a moins de microsatellites au sein des régions codantes que dans les introns et les régions 3'-UTR. Cela s'explique naturellement par une pression de sélection plus importante dans ces régions. Enfin, la distribution des SSR dans les régions 3'UTR et dans les introns ne montre pas de différence par rapport à la distribution des SSR au sein du génome total (Lawson and Zhang 2006).

▪ Rôle des SSR présents dans les régions transcrites

La présence des microsatellites va avoir un impact différent suivant qu'ils se trouvent dans les introns, dans les régions 5'- et 3'-UTR ou encore dans les régions codantes (**figure 1.6**).

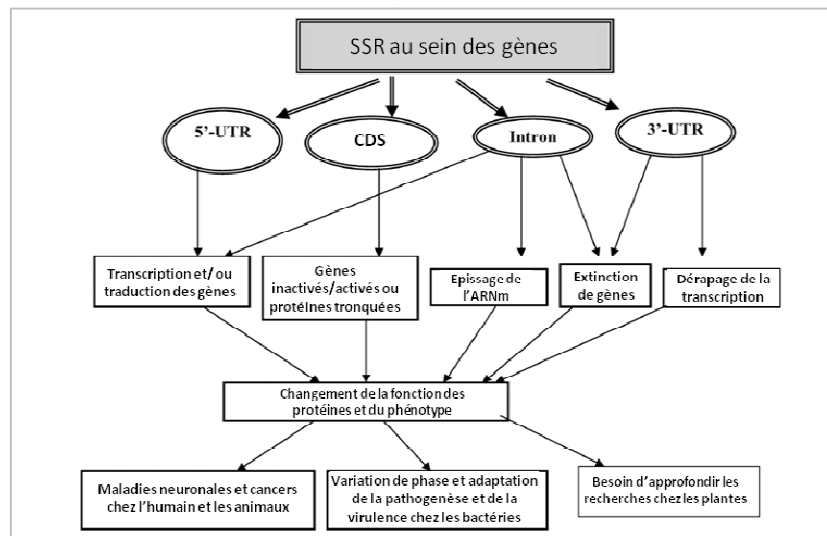


Figure 1.6 : Fonction de régulation des SSR dans les régions codantes (CDS), les 3'-et 5'-UTRs et dans les introns (Li et al. 2004)

De manière générale, la modification du nombre de répétitions d'un SSR au sein des régions transcrites peut induire un changement dans la séquence et donc dans la structure de la protéine codée. Il peut donc exister un lien entre le nombre de répétitions d'un microsatellite et un phénotype particulier (Li et al. 2004; Varshney et al. 2005). Chez le maïs, par exemple, la présence d'un SSR de type $(CCG)_n$ dans la région 5'-UTR de gène codant pour des protéines ribosomiales semble influencer la régulation de la fertilisation (Dresselhaus et al. 1999).

II.III. LES SNP (SINGLE NUCLEOTIDE POLYMORPHISM)

II.III.1. Introduction

Des erreurs peuvent se produire lors de la réplication ou même lors de la réparation de l'ADN, générant des mutations, sources de variation génétique et donc de polymorphisme. On peut classer les mutations suivant le type de changement causé (**figure 1.7**).

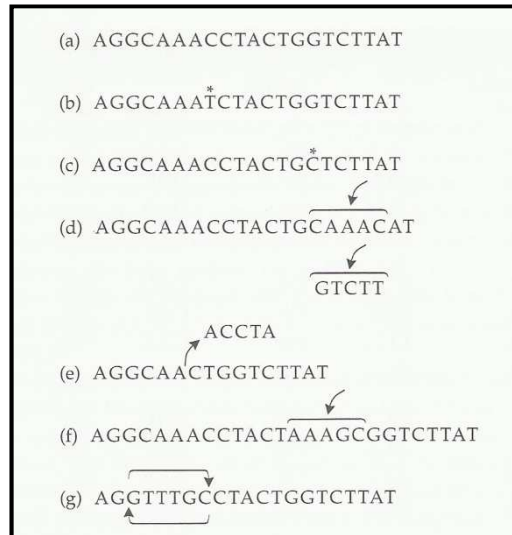


Figure 1.7 : Différents types de mutation au sein d'une séquence. (a) séquence non mutée, (b) transition, (c) transversion, (d) recombinaison, (e) délétion, (f) insertion, (g) inversion (Graur and Li 2000)

Les substitutions regroupent les permutations d'un nucléotide par un autre. Les recombinaisons décrivent l'échange d'une séquence avec une autre. Les délétions et les insertions comprennent respectivement la suppression ou l'ajout d'un ou plusieurs nucléotides dans une séquence. Enfin, les inversions sont définies comme une rotation de 180° d'un segment d'ADN double brin (Graur and Li 2000).

Le terme SNP (Single Nucleotide Polymorphism) désigne un polymorphisme de nucléotide unique, c'est-à-dire une substitution ou mutation ponctuelle. Les insertions-délétions de quelques nucléotides (indel) peuvent également être désignées par le terme SNP.

Une substitution peut être soit une transition, soit une transversion. Une transition est un changement entre A et G (entre purines) ou entre C et T (entre pyrimidines) alors qu'une transversion est un changement entre une base purine et une base pyrimidine. Il y a donc 4 types de transitions et 8 types de transversions (**figure 1.8**).

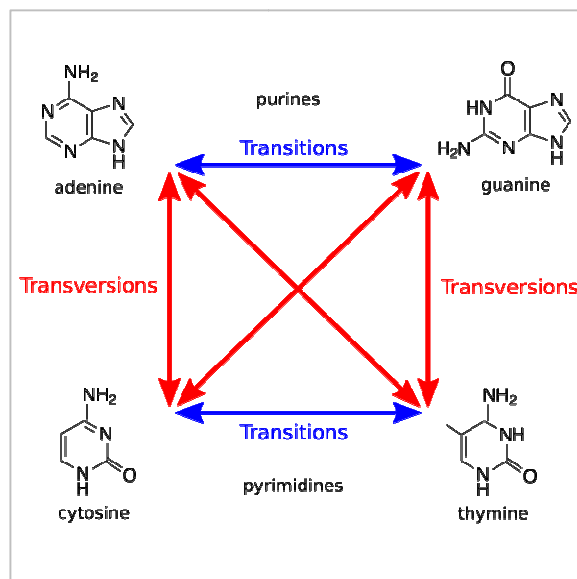


Figure 1.8 : Différentes possibilités de transitions et de transversions entre les acides nucléiques

On estime que 60 à 70 % des mutations observées au sein d'un génome sont des transitions (Graur and Li 2000). L'analyse du génome d'*A. thaliana* a montré que les transitions sont 2,4 fois plus présentes que les transversions (Ossowski et al. 2010). Cela peut s'expliquer par les mécanismes responsables des différents types de mutations. En effet, la présence d'analogues de base tels que le 5-bromouracile et la 2-aminopurine ou encore la présence d'agents mutagènes tels que l'acide nitreux, l'hydroxylamine et les agents alkylants engendre plus de transitions que de transversions (Garrett and Grisham 2000). De plus, de par la nature même du code génétique, les transitions sont moins susceptibles de changer la nature des acides aminés. En effet, lorsqu'un acide aminé correspond dans le code génétique à deux codons différents, ceux-ci possèdent au niveau de la troisième position, deux acides nucléiques appartenant à la même famille (deux purines ou deux pyrimidines). Par ce fait, une

transition pour cette position engendrera une mutation synonyme contrairement à une transversion. Ces mutations ont donc plus de chance d'être conservées au niveau des régions codantes (Gibson and Muse 2004)

II.III.2. Fréquence et influence des mutations dans les régions transcrites

Le taux de mutation chez les eucaryotes est fortement variable suivant l'espèce étudiée. Chez *Saccharomyces cerevisiae*, ce taux a été estimé à $0,33.10^{-9}$ substitutions par site nucléotidique et par génération. Chez *Mus musculus*, ce taux est de 38.10^{-9} (Lynch 2010). Chez *A. thaliana*, celui-ci est estimé à 7.10^{-9} (Ossowski et al. 2010). Au sein des régions transcrites, la fréquence ainsi que l'impact des SNP ne sont pas les mêmes suivant la zone dans laquelle on se trouve.

Le taux de mutation au sein des introns est équivalent au taux de mutation globale. Il semblerait donc que ces régions ne subissent pas de pression de sélection particulière (Graur and Li 2000). Cependant, une étude a montré que chez *A. thaliana*, la présence d'un SNP dans le septième intron du gène d'une sous-unité du protéasome induit la création d'une protéine tronquée. En effet, ce SNP crée un nouveau site d'épissage qui engendre un codon stop prématuré. La présence de SNP au sein de ces régions peut donc avoir un impact majeur sur la protéine finale et donc sur le phénotype (Guyon-Debast et al. 2010).

Les mutations présentes dans les régions codantes peuvent être définies comme synonymes (ou silencieuses) ou non-synonymes. Une mutation synonyme signifie qu'elle n'implique pas de changement dans la séquence d'acide aminé de la protéine contrairement à la mutation non-synonyme. Vu que ces mutations silencieuses ont un impact plus faible, elles subissent moins de pression de sélection. Ces mutations sont donc plus abondantes. Une analyse réalisée à l'aide de 239.172 EST de tomates a montré que la majorité des SNP présents dans les régions codantes se trouvait sur la troisième position des codons et représentait donc majoritairement des mutations silencieuses (Jimenez-Gomez and Maloof 2009). Cette observation ne signifie pas que ce type de mutation n'a pas une influence sur la protéine puisqu'une mutation silencieuse, comme celle présente dans les introns, peut engendrer la perte ou le gain d'un site d'épissage.

Puisqu'elles subissent une forte pression de sélection, les mutations non synonymes sont au final les mutations les moins fréquentes au sein des régions transcrites. Ces mutations sont soit des mutations non-sens (codon STOP prématuré), soit des mutations faux-sens (changement d'acide aminé). Si le nouvel acide aminé possède les mêmes caractéristiques physico-chimiques que le précédent, on parlera de remplacement conservatif. Dans le cas contraire on parlera de remplacement radical (Graur and Li 2000). L'impact d'une mutation non-synonyme, faux-sens et non conservative présente dans la séquence codante d'une glycosyltransferase a été étudié. Cette mutation implique le changement d'une glutamine (acide aminé polaire non chargé) en une histidine (un acide aminée polaire chargé positivement). Cette seule mutation induit un changement dans la spécificité du donneur du sucre pour cette enzyme (Kubo et al. 2004). Ce type de mutation peut donc avoir un impact important sur le phénotype de l'organisme qui le porte.

L'analyse des régions 5'- et 3'-UTR a montré un taux de substitution plus faible par rapport au taux de substitution synonyme dans les régions codantes. Cela peut s'expliquer par le fait que ces régions contiennent d'importants signaux impliqués dans la régulation de la traduction (Graur and Li 2000).

La connaissance de la nature des mutations ainsi que la zone dans laquelle elles se trouvent sont des informations importantes pour le développement de marqueurs SNP. En effet, suivant le type de région dans laquelle on se trouve (introns, exons), la fréquence des SNP sera plus ou moins élevée. De plus, les SNP présents au sein des UTR's et des régions codantes ont une plus grande probabilité d'avoir un impact sur le phénotype et donc de constituer de bons marqueurs pour la sélection.

II.IV. LES BANQUES EST : SOURCES DE POLYMORPHISME SSR ET SNP

II.IV.1. Introduction

Une banque EST est une banque de séquences provenant de la rétrotranscription de l'ARN_m. L'ARN_m provenant généralement de plusieurs individus, de nombreuses séquences se trouveront en plusieurs copies au sein de la banque.

Pour mettre en évidence des motifs SSR au sein de ces EST, les séquences présentes plusieurs fois sont regroupées en contigs et alignées. Les séquences consensus de ces contigs ainsi que les EST présentes une seule fois seront rassemblées afin d'obtenir une banque d'unigènes où les séquences ne sont présentes qu'une seule fois. Cette banque d'unigènes sera ensuite analysée à l'aide de logiciels tels que SPUTNIK (Abajian 1994) afin d'extraire les séquences possédant des motifs microsatellites.

Il est également possible d'utiliser les banques EST pour la recherche de SNP. Dans ce cas-ci, le polymorphisme est recherché au sein des séquences présentes dans un même contig. En effet, dans le cas où les banques EST proviennent des séquences transcrites de plusieurs individus, plusieurs allèles d'une même séquence peuvent être présents. C'est donc le polymorphisme présent entre ces allèles qui sera mis en évidence.

II.IV.2. Les banques EST chez les Asteraceae

Ces dernières années, les banques EST se sont développées de façon considérable. A ce jour, plus d'un million de séquences sont disponibles pour la famille des Asteraceae (**tableau 1.1**). Au sein du genre *Cichorium*, on retrouve plus de 84.000 EST.

Chez *Cichorium intybus*, plus de 50.000 EST issues de trois banques différentes ont été publiées (CGP 2012; Dauchot 2007; Legrand et al. 2007). Ces différentes banques seront décrites en détail dans le chapitre 2 consacré à l'analyse de ces données.

Tableau 1.1. Nombre de séquences EST disponibles sur le site NCBI (<http://www.ncbi.nlm.nih.gov/>) pour les différents genres de la famille des Asteraceae (mai 2012)

Genre	Nombre de séquences EST dans NCBI
<i>Helianthus</i>	323337
<i>Lactuca</i>	226709
<i>Artemisia</i>	95022
<i>Centaurea</i>	85293
<i>Cichorium</i>	84144
<i>Taraxacum</i>	57737
<i>Carthamus</i>	41584
<i>Cynara</i>	36323
<i>Barnadesia</i>	28483
<i>Gulizotia</i>	25711
<i>Zinnia</i>	20767
<i>Gerbera</i>	16997
<i>Parthenium</i>	11739
<i>Senecio</i>	10219
<i>Chrysanthemum</i>	7109
<i>Stevia</i>	5548
<i>Saussurea</i>	1897
<i>Vernonia</i>	1325
<i>Ambrosia</i>	633
autres	1267
Total:	1081844

II.IV.3. Avantages des marqueurs issus de banques EST

L'utilisation des banques EST pour le développement de marqueurs moléculaires est une stratégie peu coûteuse. En effet, les marqueurs peuvent être identifiés par l'analyse bioinformatique de banques déjà disponibles sans passer au préalable par la création de banques enrichies pour les microsatellites ou par le séquençage direct pour les SNP.

De plus, étant donné qu'il existe des banques EST pour différents groupes au sein d'une même espèce, pour différentes espèces au sein d'un même genre ainsi que pour différents genres au sein d'une même famille, il est, au final, possible de mettre en évidence du polymorphisme existant entre les différents niveaux taxonomiques.

Les mutations au sein des régions transcrites subissent plus de pression de sélection que celles présentes dans les régions intergéniques. Cela implique que, de manière générale, on s'attend à ce que le nombre de marqueurs soit moins élevé dans les EST et que le polymorphisme soit plus faible que dans les régions non-transcrites.

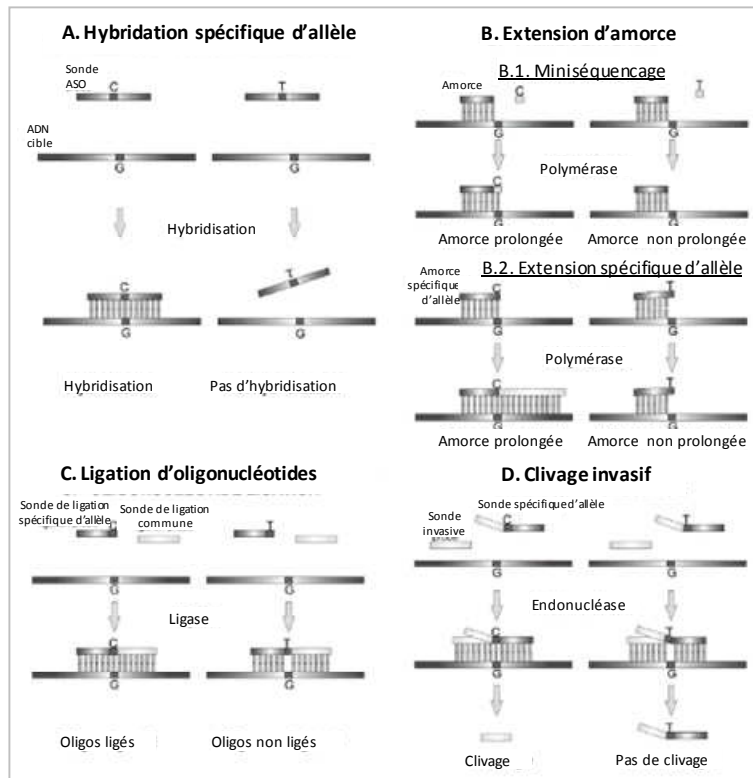


Figure 1.9 : Illustration des différentes méthodes de discrimination d'allèles. **A. L'hybridation spécifique d'allèle :** deux sondes ASO (Allelic- Specific Oligonucleotides) sont mises en présence de l'ADN cible contenant le SNP. Seule la sonde ayant une correspondance parfaite avec le brin matrice s'hybridera de manière stable. **B. Extension d'amorce. B.1. Miniséquencage :** une amorce s'hybride sur l'ADN cible juste devant la position du SNP. Cette sonde est allongée en 3' via une seule base spécifique du site polymorphe. **B.2. Extension spécifique d'allèle :** l'extrémité 3' des amorces est spécifique de chaque allèle. Dans le cas d'une hybridation parfaite, il y aura extension de l'amorce. **C. Ligation d'oligonucléotides :** deux sondes spécifiques d'allèle et une sonde commune sont nécessaires pour chaque SNP. La sonde commune s'hybride juste à côté de la sonde spécifique d'allèle. En cas d'hybridation parfaite, la ligase joindra les deux sondes. **D. Clivage invasif :** les oligonucléotides nécessaires sont une sonde invasive et des sondes spécifiques d'allèle qui s'hybrident sur l'ADN cible avec un chevauchement d'une paire de base. Quand la sonde spécifique d'allèle est complémentaire de la base polymorphe, le chevauchement en 3' de la sonde invasive forme une structure qui est reconnue et clivée par une endonucléase. Il y a alors libération de l'extrémité 5' de la sonde spécifique d'allèle (Sobrinho et al, 2005)

Les marqueurs géniques risquent donc de ne pas mettre en évidence du polymorphisme entre individus très proches. Ceux-ci pourront tout de même être utilisés pour analyser des individus plus distants. De plus, les régions transcrites étant moins variables, les marqueurs utilisés pour génotyper les sites polymorphes ont plus de chance d'être conservés et donc de pouvoir être transférés d'une espèce à l'autre.

Enfin, les marqueurs présents au sein des EST sont directement liés à un gène et donc à une fonction. Si la variabilité présente au sein de ce gène est responsable de variations phénotypiques intéressantes pour l'espèce étudiée, le marqueur lié à cette variabilité pourra être utilisé en sélection.

II.IV.1. Génotypage des marqueurs issus de banques EST

La méthode la plus utilisée pour le génotypage des SSR est l'amplification de la région du microsatellite à l'aide d'amorces spécifiques des séquences bordantes. Si les amorces sont préalablement associées à un fluorochrome, la taille des fragments peut être détectée au séquenceur. Ces marqueurs SSR sont codominants puisque tous les allèles d'un individu sont amplifiés et détectés.

Les techniques utilisées pour le génotypage des SNP sont une combinaison entre une réaction de discrimination des allèles et un mécanisme de détection des produits obtenus. Le nombre élevé de méthodes dans ces deux catégories génère une grande diversité de techniques de génotypage.

La discrimination des allèles peut être divisée en différentes catégories (Kwok 2001; Sobrino et al. 2005) (**figure 1.9**):

- **L'hybridation spécifique d'allèle** : une sonde spécifique d'allèle (ASO) vient s'hybrider sur le fragment d'ADN. Seule une hybridation parfaite sera détectée.
- **L'extension d'amorces** :
 - ✓ Miniséquençage ou incorporation d'un nucléotide spécifique d'allèle : une sonde spécifique de locus (LSO) vient s'hybrider juste à côté du SNP. L'ajout d'une base marquée de manière fluorescente permet d'identifier, suivant le fluorochrome incorporé, la nature du SNP.

- ✓ L'extension spécifique d'allèle : un ASO vient s'hybrider sur le SNP. En cas d'hybridation parfaite, l'amplification du fragment pourra se faire à l'aide d'une polymérase.
- **La ligation d'oligonucléotides** : un ASO et un LSO sont mis en présence du fragment d'ADN. S'il y a hybridation parfaite des deux sondes, une ligase pourra les relier formant ainsi un fragment spécifique.
- **Le clivage invasif** : l'hybridation parfaite d'un ASO et d'un LSO engendre un chevauchement des deux sondes qui sera clivé, libérant ainsi un rapporteur.

Les produits obtenus sont généralement associés à un fluorochrome, ce qui permet de les détecter et de les identifier soit via une migration dans un séquenceur capillaire soit via une hybridation sur damier.

II.V. MÉTHODES ALLIANT MISE EN ÉVIDENCE DU POLYMORPHISME ET GÉNOTYPAGE

II.V.1. Introduction

Certaines méthodes associées aux marqueurs moléculaires permettent la mise en évidence du polymorphisme et le génotypage de manière simultanée, ce qui représente un gain de temps considérable. De plus, ces différentes techniques permettent généralement de développer un grand nombre de marqueurs moléculaires.

La première méthode décrite ici est l'AFLP (Amplified Fragments Length Polymorphism). Malgré le fait que cette technique ait été développée il y a plus de quinze ans, elle est toujours utilisée, principalement chez les espèces non-modèles pour lesquelles on dispose de peu de bases de données génétiques.

Contrairement aux marqueurs AFLP, les marqueurs RAD (Restriction-site associated DNA) ont été développés récemment grâce aux nouvelles méthodes de séquençage. Ces marqueurs offrent de nouvelles perspectives pour le développement d'outils moléculaires chez les espèces non-modèles.

Nous parlerons également de l'Ecotilling. Dans le cadre de cette méthode, l'identification du polymorphisme et le génotypage ne sont pas simultanés mais ces deux étapes se font de façon similaire.

II.V.2. L'AFLP

- Description de la méthode

Développée par Vos et al. (Vos et al. 1995), cette technique est fondée sur la mise en évidence conjointe de polymorphisme de sites de restriction et d'hybridation d'amorces arbitraires. Cette technique utilise à la fois les enzymes de restriction et l'amplification PCR.

La mise en évidence du polymorphisme se fait en plusieurs étapes (**Figure 1.10**). L'ADN génomique est clivé par deux enzymes de restriction. Des adaptateurs de séquences connues sont ajoutés aux extrémités des fragments de restriction générant ainsi une matrice pour l'amplification. Une pré-amplification est réalisée à l'aide d'amorces de séquences complémentaires à la séquence des adaptateurs et des sites de restriction. Cette amorce est prolongée en 3' par une base arbitraire. La deuxième amplification, dite sélective, utilise des amorces identiques aux premières mais prolongées à l'extrémité 3' par deux ou trois nucléotides arbitraires. Ces amorces sélectives permettent de réduire le nombre de fragments amplifiés à une centaine. Ces fragments sont séparés et détectés par électrophorèse capillaire grâce au marquage fluorescent réalisé lors de l'amplification sélective.

- Caractéristiques des marqueurs AFLP

Les marqueurs AFLP sont anonymes au sens où l'on ne connaît pas les séquences qui leur sont associées, ce qui empêche de transférer ces marqueurs sur une autre population et de connaître les bases moléculaires du polymorphisme. Cependant, lorsqu'un marqueur AFLP montre une association statistique avec un caractère d'intérêt, il est possible, via le séquençage de la bande de migration, d'analyser la séquence afin de transformer le marqueur AFLP en STS (Sequence Tagged Site) (Dussle et al. 2002).

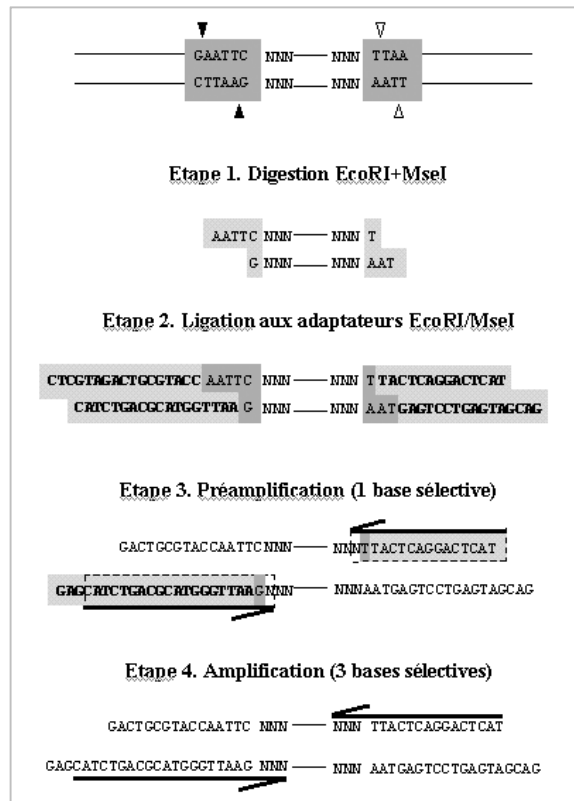


Figure 1.10 : Principe du génotypage par la méthode AFLP. Etape 1 : L'ADN génomique est digéré par deux enzymes de restriction (EcoRI et MseI). Etape 2 : Les fragments digérés par les deux enzymes sont liés à des adaptateurs spécifiques des sites de restriction. Etape 3 : Une préamplification à l'aide d'amorces spécifiques des adaptateurs et des sites de restriction est réalisée. Ces amorces possèdent en 3' une base arbitraire. Etape 4 : Une amplification sélective avec cette fois, trois bases arbitraires est effectuée. Les fragments ainsi obtenus peuvent être visualisés après migration sur séquenceur capillaire suite au marquage fluorescent réalisé lors de l'amplification sélective (Vos et al. 1995)

Les marqueurs AFLP peuvent présenter de l'homoplasie (co-migration de segments de tailles identiques mais de natures différentes). Pour une même bande de migration, nous aurons donc des fragments distincts. Plus il y a de fragments, plus il y a de risques d'avoir des fragments d'ADN différents mais de tailles identiques. La distribution de la taille des fragments AFLP n'est pas aléatoire : on observe, lors de la digestion à l'aide des enzymes EcoRI et MseI, un déficit du nombre de petits fragments et un excès du nombre de grands fragments. Ce biais engendre une augmentation de l'homoplasie par rapport à une

distribution aléatoire de la taille des fragments. Cette distribution particulière peut s'expliquer par le fait que les sites de restriction ne sont pas distribués de manière homogène dans le génome (Caballero and Quesada 2010).

La distribution des marqueurs AFLP le long du génome n'est pas aléatoire (Young et al. 1999). Certaines enzymes de restriction telles que PstI sont sensibles à la méthylation de la cytosine et coupent dans des régions peu méthylées, c'est-à-dire dans les zones transcrites. Ce type de marqueur sera donc sous-représenté dans l'hétérochromatine. Ensuite, la distribution des marqueurs AFLP sur une carte génétique est directement influencée par le taux de recombinaison qui est variable d'une région à l'autre du génome. Si le taux de recombinaison est élevé, les marqueurs seront fortement espacés. Par contre, nous observerons des regroupements de marqueurs dans les zones où le taux de recombinaison est faible. Enfin, le pourcentage local en GC, plus élevé dans les régions riches en gènes, influence indirectement la distribution des marqueurs puisqu'elle influence la répartition des sites de restriction.

- Utilisation des marqueurs AFLP chez la chicorée

L'AFLP présente l'avantage de fournir un grand nombre de marqueurs en peu de temps, sans recherche préalable de polymorphisme. L'AFLP est donc fréquemment utilisé chez des espèces non-modèles où les ressources génétiques sont rares, comme chez la chicorée : avant 2006, aucune ressource génétique n'était disponible pour cette espèce. Les premières applications des marqueurs AFLP chez la chicorée ont été l'identification, l'évaluation de la diversité génétique et l'analyse phylogénique des différents cultivars (Koch and Jung 1997; Van Cutsem et al. 2003; Van Stallen et al. 2000) et des différentes espèces du genre *Cichorium* (Kiers et al. 2000).

En 1997, Koch a étudié 17 variétés de chicorée industrielle à l'aide des marqueurs AFLP et de marqueurs RAPD (Random Amplification of Polymorphic DNA). Il en conclut que ces variétés étaient phylogénétiquement très proches et qu'il existait une grande diversité génétique intravariétale. En 2000, Van Stallen a identifié les profils génétiques issus de marqueurs AFLP et RAPD spécifiques à chaque variété de chicorée witloof étudiée. Une faible diversité génétique au sein de ces variétés a été observée. Van Cutsem et al. en 2003 sont parvenus à distinguer à l'aide de marqueurs AFLP, les individus

sauvages de chicorée par rapport aux individus cultivés. Le niveau de diversité génétique observé entre ces deux groupes semblait identique. Enfin, en 2000, Kiers a réalisé une étude phylogénétique des différentes espèces du genre *Cichorium* à l'aide de marqueurs AFLP. Malgré l'analyse phylogénétique concluante, il n'a pas été permis de trouver des marqueurs AFLP spécifiques à chaque espèce.

Ces études ont permis de développer des outils moléculaires utiles pour le sélectionneur. Tout d'abord, ces marqueurs permettent l'identification sans équivoque de certaines variétés. Ensuite, les distances génétiques entre différentes variétés et différentes espèces peuvent être calculées. Cela permet d'avoir une meilleure connaissance des relations phylogénétiques afin d'identifier les variétés génétiquement éloignées en vue de nouveaux croisements. Enfin, l'étude de la diversité génétique permet de savoir au sein de quelle variété il y a le plus de ressources génétiques pour la création de nouvelles populations.

Deux cartes génétiques de *Cichorium intybus* ont également été réalisées à l'aide de marqueurs AFLP (De Simone et al. 1997; Van Stalle et al. 2003). L'AFLP permet en effet d'obtenir un grand nombre de marqueurs en peu de temps, ce qui est donc très utile pour la saturation de cartes génétiques. Nous détaillerons ces études dans le chapitre 4 consacré à la cartographie.

II.V.3. Les marqueurs RAD

Le séquençage nouvelle génération (NGS) permet d'obtenir rapidement un très grand nombre de données pouvant être utilisées pour la mise en évidence de polymorphisme (Egan et al. 2012).

Le séquençage multiplex permet de séquencer et de génotyper plusieurs individus en même temps à l'aide de milliers de marqueurs. Pour ce faire, l'ADN génomique est fragmenté de manière aléatoire et chaque fragment est couplé à un « code-barres » de quelques nucléotides qui est identique pour tous les fragments d'un même individu. La présence de ce code-barres permet de séquencer les fragments de tous les individus de manière simultanée. Cette méthode est adaptée pour les génomes de petite taille comme le riz (500 Mb) (Huang et al. 2009) ou encore pour l'analyse de l'ADN mitochondrial et chloroplastique (Cronn et al. 2008). Cependant, lorsque la taille du

génomique est plus élevée, le recouvrement des séquences est moins fréquent, ce qui rend cette méthode moins performante.

Afin de contrecarrer ce problème, il est possible de diminuer la complexité du génome c'est-à-dire d'éliminer certaines régions possédant des caractéristiques particulières à l'aide d'enzymes de restriction. Dans ce cas, seules les régions se trouvant aux alentours des sites de restrictions digérées seront séquencées. Suite à l'analyse bioinformatique des données, on peut mettre en évidence les SNP au sein des alignements mais également le polymorphisme présent au niveau des sites de restriction. Les marqueurs obtenus par ce principe sont appelés marqueurs RAD pour Restriction-site Associated DNA (Baird et al. 2008).

La méthode dite de Genotyping-by-Sequencing (GBS) permet le développement de ces marqueurs. La **figure 1.11** décrit en détail le principe du GBS. Au final, cette technique permet, via la réduction de la complexité du génome par l'utilisation d'enzymes de restriction, de développer des marqueurs codominants de type SNP et des marqueurs dominants issus du polymorphisme au sein des sites de restriction (Elshire et al. 2011).

Un des facteurs déterminants de cette méthode est le choix de l'enzyme de restriction. En effet, celle-ci a un impact direct sur le niveau de réduction de la complexité du génome. Elshire a utilisé le GBS pour l'analyse du génome du maïs. L'enzyme de restriction ApeKI a été choisie pour plusieurs raisons. Cette enzyme coupe un site composé de cinq bases. La taille du site influence directement la complexité du génome retenu. En effet, un site de restriction de grande taille est moins fréquent dans le génome et donc l'enzyme générera moins de fragments. Cette enzyme a également été sélectionnée car son site est peu présent dans les différentes classes de retrotransposons chez le maïs. Cela permet donc d'éliminer en partie ces séquences répétées. Enfin, cette enzyme est sensible à la méthylation, ce qui signifie que les régions fortement méthylées seront peu représentées dans les fragments analysés. Cet exemple montre l'importance du choix de l'enzyme de restriction pour ce type d'analyse. Suivant l'espèce, il sera nécessaire d'adapter les critères de sélection (Elshire et al. 2011).

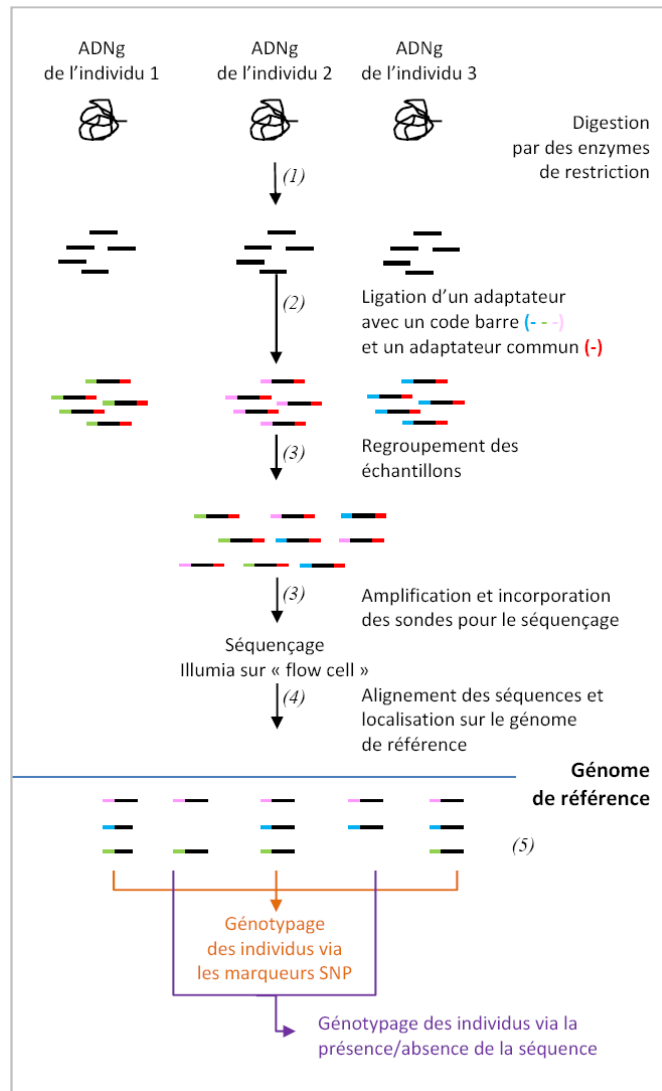


Figure 1.11: Principe du Genotyping-by-Sequencing (GBS). (1) L'ADNg de plusieurs individus est digéré à l'aide d'une enzyme de restriction. (2) Les fragments obtenus sont ligés à un adaptateur possédant un code-barres spécifique de chaque individu et un adaptateur commun pour tous les fragments digérés. (3) Tous les échantillons sont regroupés et une amplification est réalisée afin d'incorporer en 5' et en 3' les séquences nécessaires pour le séquençage Illumina sur « flow cell ». (4) Les séquences obtenues (86 pb) sont alignées et positionnées sur un génome de référence. (5) Pour les séquences présentes chez tous les individus, l'alignement permettra de mettre en évidence du polymorphisme de type SNP. Pour les séquences présentant du polymorphisme au sein des sites de restriction, le génotypage se fera via la présence/absence du fragment chez les différents individus.

Le développement de marqueurs RAD via le GBS permet pour un prix aujourd'hui abordable de géotyper en une seule fois des centaines d'individus à l'aide de plusieurs milliers de marqueurs.

II.V.4. L'Ecotilling

L'Ecotilling est une méthode qui permet de mettre en évidence du polymorphisme via la formation et la digestion d'hétéroduplexes obtenus grâce à l'hybridation non-parfaite de séquences polymorphes. Cette méthode permet également de réaliser le géotypage d'individus en présence d'un ADN de référence. La **figure 1.12** décrit en détail ces deux aspects (détection du polymorphisme et géotypage) de l'Ecotilling. Le polymorphisme détecté et géotypé peut être de type SSR, SNP ou encore Indel (Comai et al. 2004). Cette méthode a l'avantage de détecter du polymorphisme entre plusieurs individus en une seule fois, sans définition préalable de sondes spécifiques pour chaque locus polymorphe.

L'Ecotilling a été utilisé chez la chicorée dans le but de mettre en évidence et d'analyser le polymorphisme au sein de séquences codantes pour des enzymes impliquées dans le métabolisme de l'inuline. Suite à cette étude, une insertion/délétion a pu être mise en évidence au sein d'une FEH (Fructane Exohydrolase), une enzyme responsable de la dégradation de l'inuline (Dauchot 2007). L'étude plus approfondie de cette séquence a permis d'identifier une association statistique entre le polymorphisme nucléotidique présent au sein d'un groupe de trois FEH et le degré de polymérisation de l'inuline après exposition au froid (N. Dauchot, communication personnelle). Cette étude a donc permis de développer un marqueur moléculaire pour la sélection d'individus peu sensibles à la dépolymérisation.

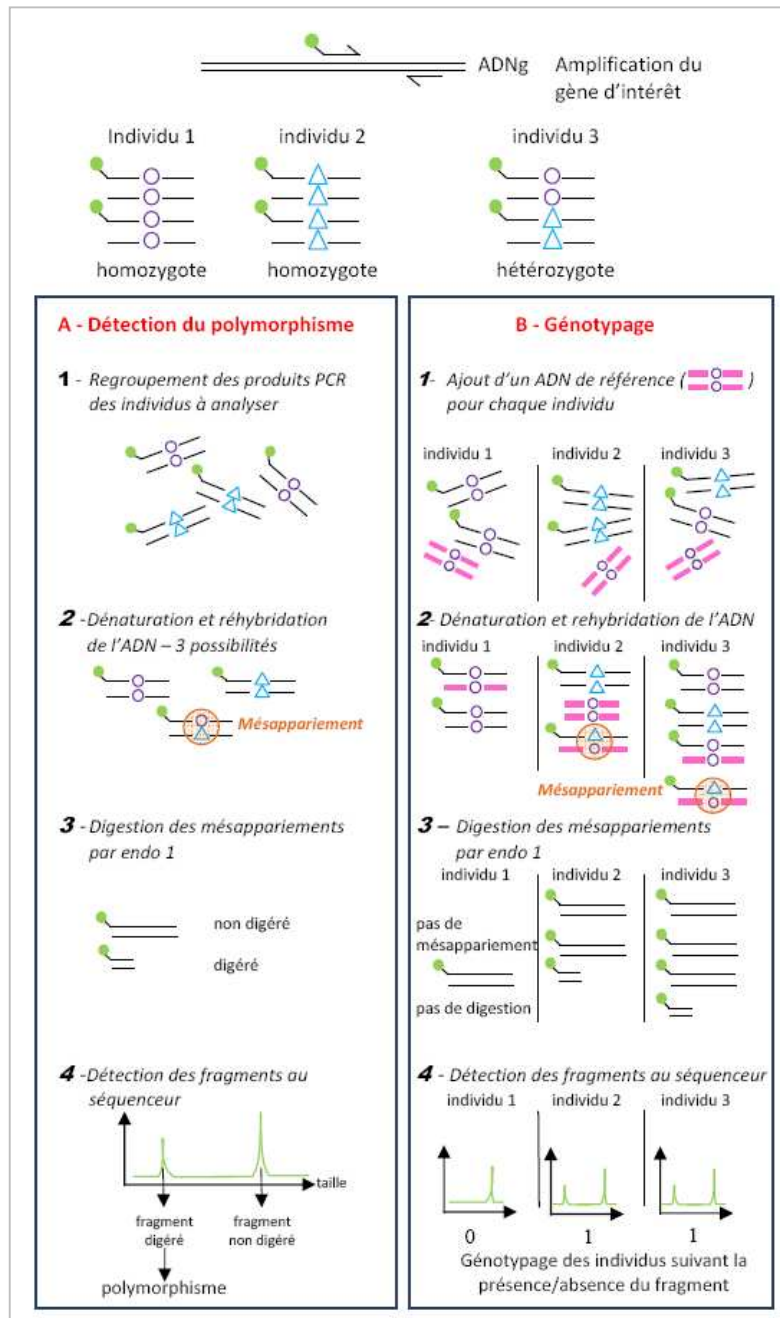


Figure1.12 : La méthode d'Ecotilling. Le gène d'intérêt doit être préalablement amplifié. Le produit d'amplification ainsi formé est marqué de manière fluorescente grâce à l'utilisation d'une amorce marquée. **A)** Utilisation de l'Ecotilling pour la mise en évidence de polymorphisme. (1) Les produits d'amplification provenant de plusieurs individus sont regroupés dans un seul échantillon. (2) Suite à la dénaturation et à la réhybridation des produits d'amplification, il y a formation de mésappariements pour les brins d'ADN provenant d'allèles différents. (3) L'utilisation d'une endonuclase permet de couper les fragments d'ADN comportant des mésappariements. (4) La migration sur séquenceur capillaire permet d'identifier la taille des différents fragments obtenus. Le plus grand fragment correspond au produit d'amplification non digéré. Les fragments de tailles intermédiaires correspondent à la position du polymorphisme présent entre les individus étudiés. **B)** Utilisation de l'Ecotilling pour le génotypage. (1) Pour chaque individu, un ADN de référence correspondant à la séquence étudiée et dont on connaît le génotype est ajouté à l'échantillon. (2) Suite à la dénaturation et à la réhybridation de l'ADN, il y aura formation de mésappariements si l'individu est hétérozygote (individu 3) ou s'il possède des allèles différents par rapport à la référence (individu 2). (3) L'action de l'endonucléase va permettre de cliver l'ADN au niveau des mésappariements. (4) La migration des échantillons sur un séquenceur capillaire permet d'identifier la taille des différents fragments obtenus après digestion. Le génotypage des individus se fait suivant la présence/absence du fragment. Ces marqueurs sont donc via cette démarche, de type dominant

III. LA CARTOGRAPHIE GÉNÉTIQUE

III.I. INTRODUCTION

Une carte génétique est une représentation du génome sur laquelle sont placés des marqueurs dont on connaît les positions relatives sur les groupes de liaison représentant les chromosomes. Ces positions sont déterminées en analysant la façon dont les marqueurs ségrègent parmi les descendants d'un croisement (Lorieux 1994b). Lors de ce travail, le polymorphisme mis en évidence chez la chicorée a été utilisé pour construire une carte génétique.

La cartographie nécessite donc une population de ségrégation, issue d'événements méiotiques où surviennent des recombinaisons entre chromosomes homologues (crossing-over). Le taux de recombinaison entre deux loci marqueurs est d'autant plus grand que ceux-ci sont éloignés sur le chromosome. Le calcul de la fréquence de recombinaison au sein de la descendance du croisement permet donc d'estimer la distance génétique séparant ces deux loci.

III.II. TYPES DE POPULATION

Ils existe plusieurs types de populations habituellement utilisés pour la construction de cartes génétiques (Samouelian et al. 2009).

Les individus d'une génération F2 sont issus d'un croisement de deux parents supposés homozygotes, suivi d'une autofécondation d'un individu (**figure 1.13.A**). Ce type de population est très hétérogène, ce qui la rend éphémère. Il est donc nécessaire d'isoler et de conserver l'ADN des différents individus.

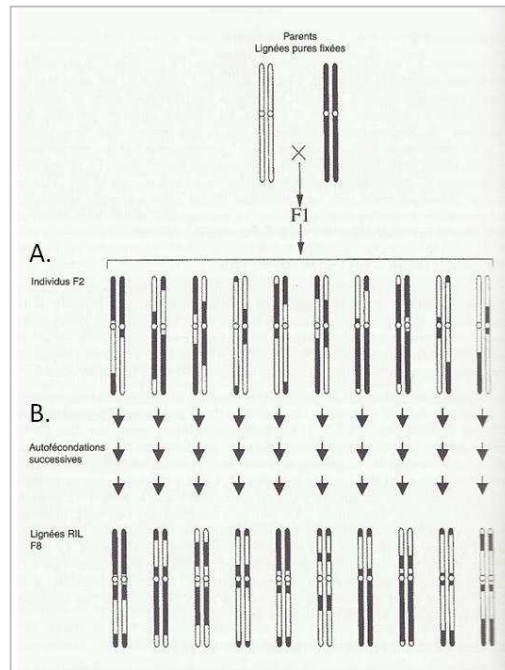


Figure 1.13 : Obtention d'une population F2 et de lignées recombinantes RIL. Chaque génome est symbolisé par une paire de chromosomes. **A. Obtention d'une population F2 :** deux parents provenant de lignées fixées sont croisés. Une autofécondation est ensuite réalisée sur un individu de la F1 générant ainsi la population F2. Cette population est en ségrégation et chaque individu est différent. **B. Obtention de lignées recombinantes (RIL) :** les individus de la F2 subissent une succession d'autofécondations. A la génération F8, les lignées sont pratiquement homozygotes et représentent la fixation d'une combinaison unique et aléatoire des génomes parentaux (Samouelian et al. 2009)

Les individus d'un rétrocroisement (backcross) proviennent du croisement entre un individu de la génération F1 avec un des parents homozygotes. Ce type de population est souvent utilisé pour l'introduction de nouveaux caractères au sein d'une variété. Les rétrocroisements permettent de fixer les événements de recombinaison ayant lieu à la génération F1. De plus, l'analyse des résultats est plus simple vu que pour deux locus donnés, il n'y aura que quatre classes différentes d'individus.

Les lignées recombinantes consanguines ou pures (RIL, Recombinant Inbred Line) sont construites de la même façon qu'une population F2. Cependant, cette génération F2 est suivie d'environ six autofécondations. La lignée obtenue est donc quasi homozygote et donc facilement conservée (**figure 1.13.B**).

Les plantes haploïdes peuvent être obtenues via la culture *in vitro* d'anthers. Ensuite, en présence de colchicine, le nombre de chromosomes est doublé. Cela permet donc d'obtenir des plantes haploïdes doublées homozygotes et fertiles. Si les individus haploïdes proviennent d'une génération F1, les haploïdes doublés auront des génotypes variables. Cette technique peut être utilisée pour la création de descendants mais également pour l'obtention de parents homozygotes.

III.III. PRINCIPE DE LA CARTOGRAPHIE GÉNÉTIQUE

La construction d'une carte génétique via les différents logiciels se fait en trois étapes : premièrement, l'assignation des marqueurs à un groupe de liaison, deuxièmement l'ordonnancement de ces marqueurs au sein de chaque LG (groupe de liaison) et troisièmement, l'établissement des distances génétiques entre les marqueurs (Cheema and Dicks 2009)

III.III.1. Assignation des marqueurs à un groupe de liaison

Lors de cette étape, deux valeurs sont calculées pour tous les couples de marqueurs : l'estimation de la fréquence de recombinaison r est calculée grâce à la méthode du maximum de vraisemblance et la détection de la liaison est estimée à l'aide du LOD score.

Le principe du maximum de vraisemblance est le suivant : on écrit la vraisemblance d'obtenir les effectifs attendus des différents génotypes

en fonction de la fréquence de recombinaison (r) et des effectifs observés. Ensuite, suivant cette équation, on détermine la valeur de r pour laquelle la vraisemblance est maximale. Suivant le type de population et le type de marqueurs (dominants, co-dominants) utilisés, la résolution de l'équation sera différente. Dans le cas d'un rétrocroisement, l'estimation de r revient à sa plus simple expression c'est-à-dire :

$$\hat{r} = \frac{b+c}{n}$$

Où r est l'estimation du taux de recombinaison entre deux loci, b et c sont les effectifs observés pour les génotypes recombinants et n est le nombre total d'individus dans la population.

Les équations permettant d'estimer r dans une population F2 à l'aide de marqueurs dominants et co-dominants sont plus complexes. Cependant, le principe reste identique puisque la valeur de r dépend dans tous les cas des effectifs observés pour les différents génotypes, ainsi que de la taille de la population (Lorieux 1994a).

Afin de tester la liaison entre deux loci, le LOD score, c'est-à-dire un rapport de vraisemblance, est calculé d'après la formule suivante :

$$\text{LOD} = \log_{10} \{V(r)/V(r_0)\}$$

Où $V(r)$ est la vraisemblance calculée avec la valeur la plus probable de r et $V(r_0)$ est la vraisemblance calculée pour $r = 0,5$. Cette valeur représente donc le rapport entre la probabilité d'obtenir les résultats observés sous l'hypothèse d'une liaison avec un taux de recombinaison r et la probabilité sous l'hypothèse d'indépendance. Ce rapport étant logarithmique, un LOD score de 3 signifie que la liaison des marqueurs est 1000 fois plus probable que l'indépendance (de Vienne 1998).

La détermination des valeurs seuils du taux de recombinaison ainsi que du LOD score vont permettre de définir les marqueurs se trouvant dans les mêmes groupes de liaison.

III.III.2. Etablissement de l'ordre des marqueurs dans chaque LG

L'ordre des loci au sein de chaque LG doit être défini. Il n'est pas possible d'établir l'ordre des marqueurs en une seule fois quand ceux-ci sont trop nombreux. En effet, pour un nombre m de marqueurs, il existe $m!/2$ ordres possibles soit 2520 pour 7 marqueurs. Il faut donc d'abord ordonner les marqueurs les plus informatifs via une analyse en trois points. Le but de cette étape intermédiaire est de positionner des marqueurs dont l'ordre est soutenu par une probabilité élevée. Pour pouvoir répondre à ce critère, le groupe sélectionné (hautement informatif) doit contenir des marqueurs pour lesquels il manque peu de données génotypiques et situés à une distance suffisante les uns des autres afin d'être repartis sur tout le génome (Lincoln et al. 1993). Les marqueurs restants sont ensuite ajoutés via une analyse multipoints. Cette stratégie en deux étapes permet de diminuer fortement le nombre de calculs.

III.III.3. Détermination des distances génétiques

Une fois l'ordre établi au sein d'un LG, la distance entre marqueurs peut être calculée. Cette valeur est exprimée en centiMorgan (cM). Le centiMorgan est fonction du taux de recombinaison observé entre deux marqueurs mais il tient compte également du fait que les recombinaisons paires (doubles crossings-over) ne peuvent être détectées. Il existe plusieurs types de fonctions pour transformer la fréquence de recombinaison en centiMorgan. Les fonctions de Haldane et de Kosambi sont de loin les plus utilisées :

- La fonction de Haldane : celle-ci se base sur l'hypothèse qu'il y a absence d'interférence génétique. La probabilité d'occurrence d'un crossing-over dans un segment quelconque est indépendante de l'occurrence d'un crossing-over dans un autre fragment, même très proche du premier.
- La fonction de Kosambi : cette fonction prend en compte l'occurrence c'est-à-dire que la probabilité d'occurrence d'un crossing-over dans un segment quelconque est dépendante de l'occurrence d'un crossing-over dans un autre segment.

Ces fonctions permettent donc de faire le lien entre l'estimation du taux de recombinaison et la distance séparant deux marqueurs. Un cM

correspond au final non pas à 1% de recombinaison entre deux loci mais bien à la distance séparant deux loci entre lesquels l'espérance du nombre de crossing-over est de un centième.

Plusieurs logiciels ont été développés afin de permettre de réaliser les différentes étapes pour la construction d'une carte génétique. Parmi ceux-ci, on retrouve MapMaker (Lander et al. 1987), JoinMap (Stam 1993), CarthaGene (Schiex and Gaspin 1997), RECORD (Van Os et al. 2005), AntMap (Iwata and Ninomiya 2006), MST_{MAP} (Wu et al. 2008) ou encore THREaD Mapper (Cheema et al. 2010). La démarche utilisée pour la cartographie est comparable d'un programme à l'autre (Cheema and Dicks 2009). Le choix de l'outil va dépendre de différents critères tels que le type de croisement pris en compte, l'interface ou encore les fonctionnalités mises en œuvre pour vérifier la qualité de la carte.

III.IV. LA CARTOGRAPHIE CHEZ LA CHICORÉE ET CHEZ LA LAITUE

A ce jour, trois cartes génétiques pour le genre *Cichorium* ont été publiées. Tout d'abord, une carte issue d'un croisement entre *C. intybus* et *C. endiva* a été utilisée pour la création d'une carte génétique constituée de marqueurs RAPD (Random Amplification of Polymorphic DNA), AFLP et SAMPL (Selective Amplification of Microsatellite Polymorphic Loci) (De Simone et al. 1997). En 2003, une carte génétique provenant d'un croisement intraspécifique a été réalisée à l'aide de 129 marqueurs RAPD (Van Stallen et al. 2003). Enfin, en 2010, une carte consensus de *C. intybus* a été construite avec 431 SSR et 41 STS (Sequence-Tagged Sites) (Cadalen et al. 2010).

Depuis 2000, l'Université de Californie à Davis aux USA a mis en place le « Compositae Genome Project » (CGP) qui a pour but de développer un catalogue de séquences génomiques ainsi que des cartes génétiques détaillées pour les espèces appartenant à la famille des Asteraceae. Un des genres les plus proches de *Cichorium* et largement étudié dans le cadre de ce consortium est le genre *Lactuca*. Une dizaine de cartes génétiques sont disponibles sur le site du CGP (<http://compgenomics.ucdavis.edu/>) pour cette espèce. Sept cartes composées de marqueurs AFLP, RFLP, SSR et RAPD sont ainsi disponibles. A partir de celles-ci, une carte consensus de 2.744 marqueurs a été construite (Truco et al. 2007). De plus, trois cartes composées de plus de 450 EST dont les séquences sont disponibles

ont également été mises en ligne. Enfin, une carte génétique construite à l'aide de 294 gènes candidats associés à la résistance aux pathogènes est également publiée et disponible sur le site (McHale et al. 2009).

III.V. UTILISATION DES CARTES GÉNÉTIQUES

III.V.1. Localisation et caractérisation de QTL

De nombreux caractères d'intérêt agronomique présentent une distribution continue. Ce sont des caractères quantitatifs tels que la taille, le poids, la production de sucre, la sensibilité à la montaison, la forme des fruits,... Ces traits sont majoritairement polygéniques, c'est-à-dire contrôlés par plusieurs gènes. Les loci responsables de la variation continue d'un caractère sont appelés QTL (Quantitatif Trait Loci) (Samouelian et al. 2009).

Les marqueurs moléculaires permettent de localiser les QTL sur une carte génétique. En effet, au sein d'une population, il est parfois possible d'observer une différence significative pour la distribution du trait quantitatif entre les deux homozygotes d'un locus marqueur. Dans ce cas, on peut donc supposer que le locus marqueur et le QTL sont proches l'un de l'autre sur la carte génétique (**figure 1.14**). En outre, la présence de marqueurs au sein de QTL permet d'utiliser ceux-ci pour le suivi des traits quantitatifs.

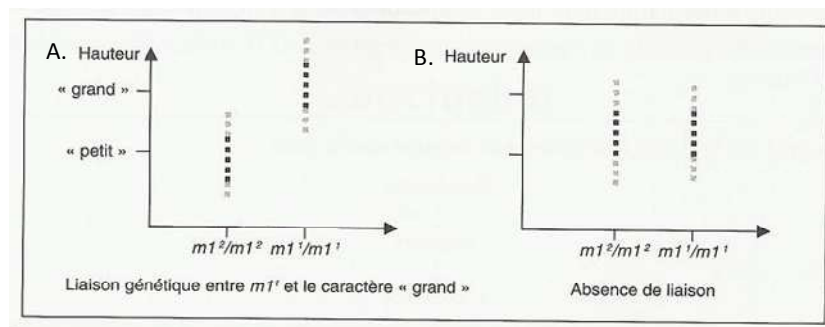


Figure 1.14 : Association entre un marqueur $m1$ et un trait quantitatif responsable du contrôle de la taille de la plante. La taille des plantes a été mesurée et classée en fonction du génotype de l'individu pour un marqueur donné. **A.** Il existe une association statistique entre l'allèle $m1^1$ et le caractère « grand ». Le marqueur $m1$ et le QTL sont donc liés génétiquement. **B.** La distribution des caractères ne semble pas associée au marqueur $m1$. Il y a donc absence de liaison (de Vienne, 1997, Samouelian et al. 2009)

A ce jour, la cartographie des QTL est toujours largement utilisée pour l'étude d'espèces cultivées (Cao et al. 2012; Ingala et al. 2012; Madrid et al. 2012). Par exemple, en 2012, Esteras et al. ont construit une carte génétique de *Cucurbita pepo* (les courges d'été) comprenant 304 marqueurs SNP et 11 marqueurs SSR. Ils ont localisé sur celle-ci des QTL contrôlant la floraison ainsi que la couleur et la forme des fruits. Les marqueurs moléculaires localisés dans ces QTL ont été utilisés pour valider ceux-ci au sein de rétrocroisements. Chez la chicorée witloof, plusieurs QTL associés à la production, au forçage, ainsi qu'à la coloration ont été identifiés (Van Stallen et al. 2003). Chez la laitue, plusieurs études ont permis de développer des QTL en relation avec la résistance aux pathogènes (McHale et al. 2009), la longévité des graines suivant les conditions de stockage (Schwember and Bradford 2010), la germination (Argyris et al. 2011) ou encore le contenu en chlorophylle et la capacité d'absorbance des radicaux libres dérivés de l'oxygène (Hayashi et al. 2012).

L'analyse conjointe de QTL et de marqueurs moléculaires pour le clonage positionnel permet d'aller beaucoup plus loin dans la caractérisation des gènes influençant les caractères quantitatifs. Par exemple, Cai et al. en 2012 ont localisé sur la carte génétique du riz un QTL impliqué dans la date d'épiaison. Ce QTL recouvre environ 30 cM du génome. L'ajout de marqueurs au sein de cette région a permis d'identifier ceux ne présentant pas ou quasi pas de recombinaison avec le QTL et donc très proches de celui-ci. La position de ces marqueurs sur la séquence génomique du riz a permis de délimiter une région de 299kb. Cette zone comportant une dizaine de marqueurs a pu être restreinte à 31.8 kb grâce à la technique dite de cartographie par substitution (substitution mapping). Les BAC associés à cette région ont été identifiés grâce aux marqueurs bordants. Le séquençage a permis de mettre en évidence un gène candidat (un facteur de transcription liant la boîte CCAAT)pouvant être responsable de la variation du trait quantitatif (Cai et al. 2012)..

III.V.2. La sélection assistée par marqueur

Quand un caractère agronomique intéressant a été identifié et localisé sur le génome, on peut l'introduire au sein de variétés cultivées : c'est l'introgression. La variété possédant le caractère d'intérêt est croisée avec la variété cultivée et plusieurs rétrocroisements sont ensuite réalisés entre la F1 et le parent cultivé. A chaque cycle, les individus

possédant le caractère d'intérêt et ressemblant au parent cultivé sont sélectionnés.

En sélection classique, 7 à 10 rétrocroisements sont nécessaires, ce qui peut prendre plusieurs années suivant la durée du cycle de reproduction. L'utilisation de marqueurs moléculaires liés et non-liés au caractère d'intérêt permet d'analyser les individus beaucoup plus rapidement, et également de diminuer le nombre de rétrocroisements. En effet, le génotypage, contrairement au phénotypage, peut être réalisé à un stade précoce de développement de la plante. De plus, l'utilisation de marqueurs permet de contrôler la taille du fragment chromosomique « introgressé » au sein de l'espèce cultivée.

III.V.3. L'assemblage des génomes

Avec le séquençage à haut débit, les cartes génétiques peuvent paraître désuètes mais constituent toujours un outil très utile pour établir l'ordre des données séquencées, et donc pour l'assemblage des génomes.

Chez la pastèque (*Citrullus lanatus*), une carte génétique comportant 953 marqueurs répartis sur 11 groupes de liaison a été construite. Grâce à celle-ci, 234 « échafaudages » issus de l'assemblage de séquences ont pu être localisés sur le génome, et chaque groupe de liaison a été associé à un chromosome à l'aide de la méthode FISH (Fluorescent In Situ Hybridation) via 11 BAC spécifiques de chaque groupe de liaison (Ren et al. 2012).

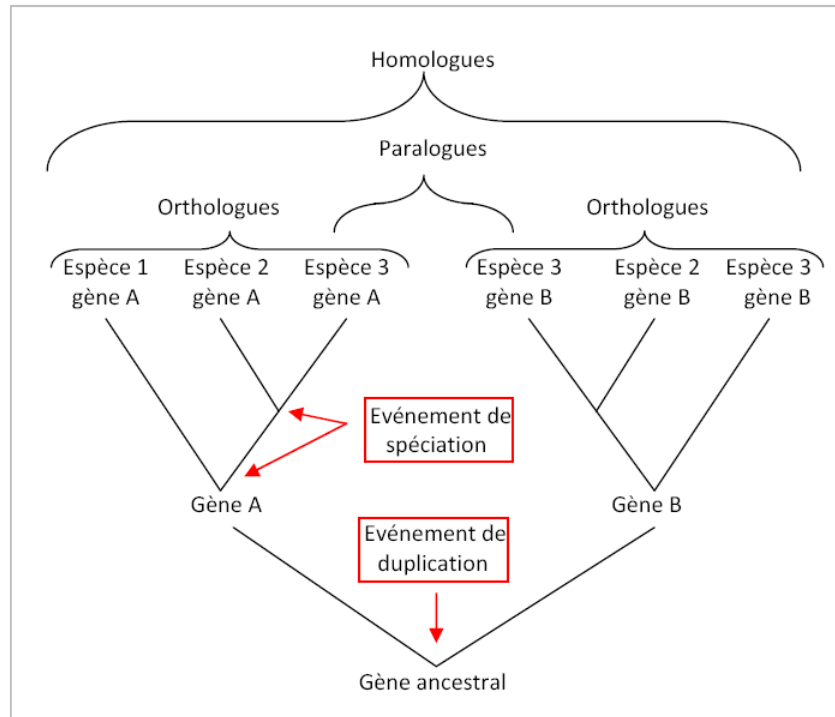


Figure 1.15: Relations possibles entre séquences homologues. Séquences homologues : séquences provenant d'une séquence ancestrale commune. Séquences paralogues : séquences homologues dérivées d'un événement de duplication. Séquences orthologues : séquences homologues dérivées d'un événement de spéciation

IV. LA CARTOGRAPHIE COMPARÉE

IV.1. INTRODUCTION

La cartographie comparée est fondée sur la localisation de séquences orthologues sur le génome de deux ou plusieurs espèces. Les séquences orthologues sont des séquences homologues dérivées d'un événement de spéciation à partir d'une seule séquence ancestrale commune. Les séquences paralogues sont des séquences homologues dérivées d'un événement de duplication (**figure 1.15**). Les gènes orthologues ont généralement des fonctions similaires chez des espèces étroitement liées (Kuzniar et al. 2008). Le concept d'orthologie est souvent mal utilisé par référence à des gènes ayant une fonction équivalente dans différentes espèces alors que c'est un concept strictement évolutif (Fitch 2000).

A ce jour, il est possible d'identifier des paires d'orthologues entre deux espèces grâce à l'existence de nombreuses séquences issues de banques EST et de séquences complètes de génome. L'analyse BLAST de ces données permet de mettre en évidence, entre deux espèces, les paires de séquences possédant la plus grande homologie.

Une fois les séquences orthologues localisées sur les génomes, l'analyse de la synténie peut être réalisée. La synténie est définie comme la conservation de l'ordre des gènes ou des marqueurs entre les cartographies ou les génomes d'espèces différentes. Le terme synténie englobe donc, généralement, le concept de colinéarité. Deux ordres de grandeur peuvent être définis. La macrosynténie s'intéresse à la conservation à l'échelle de grands blocs chromosomiques ou de grandes distances génétiques. La microsynténie fait appel à la séquence nucléotidique et se base sur des régions plus courtes (Eckardt 2001; Morgot-Gaudry and Briat 2004).

De nombreux réarrangements intra- et inter-chromosomiques s'accumulent au sein d'un génome. Plus deux espèces sont éloignées, plus on s'attend à observer un faible niveau de synténie et inversement. Ce type d'analyse permet d'émettre des hypothèses sur l'histoire évolutive des génomes jusqu'à leur ancêtre commun (Nadeau and Sankoff 1998). De manière plus appliquée, la présence

de synténie permet également de transférer des informations génétiques d'une espèce à l'autre.

IV.II. DÉMARCHE

IV.II.1. Choix du niveau taxonomique étudié

La cartographie comparée, suivant le but à atteindre, peut être réalisée entre espèces de même genre, de même tribu, de même famille ou entre différentes familles. Plus deux espèces sont phylogénétiquement éloignées, plus les séquences orthologues sont divergentes et donc difficiles à mettre en évidence. En outre, la présence d'événements de polyploïdisation chez une des deux espèces augmente la difficulté de distinguer les orthologues des paralogues (Nadeau and Sankoff 1998; Spannagl et al. 2011).

La construction de génome ancestral est une démarche qui permet d'étudier des génomes fortement divergents comme ceux d'*A. thaliana* et du genre *Brassica*. *A. thaliana*, de la famille des Brassicaceae et de la tribu des Camelinaeae, a été choisie comme espèce modèle chez les dicotylédones en raison de son petit génome (157 Mb), de son nombre peu élevé de chromosomes ($n=5$), de sa faible taille et de son cycle de développement court. En termes de cartographie comparée, cette espèce n'est toutefois pas idéale : *A. thaliana* a subi trois duplications de génome suivies de nombreux remaniements chromosomiques. Elle possède ainsi un nombre particulier ($n=5$) de chromosomes par rapport aux autres espèces de cette famille (37% des Brassicaceae ont un nombre de chromosomes $n=8$). Afin de faciliter la comparaison des génomes, un caryotype ancestral basé sur la cartographie comparée de plusieurs espèces de la tribu des Camelinaeae et des Brassicaceae a été construit. Cette démarche permet, au final, de comparer le génome d'espèces éloignées comme *A. thaliana* et *Brassica* (Schranz et al. 2006).

La recherche de gènes dupliqués dans les banques EST de 18 espèces d'Asteraceae a mis en évidence trois duplications totales du génome. Le premier, commun à toutes les espèces, a eu lieu près de l'origine de la famille, juste avant la radiation en tribus. Les deux autres événements de duplication ont eu lieu à la base des tribus Mutisieae et Heliantheae (**figure 1.16**) (Barker et al. 2008). Les problèmes engendrés par les événements de duplication totale ne devraient

donc pas être rencontrés pour la cartographie comparée entre espèces appartenant à la tribu des Cichorioideae dont fait partie la chicorée.

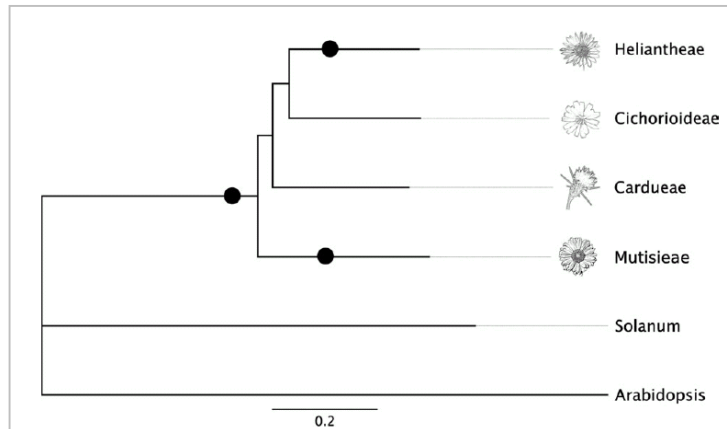


Figure 1.16 : Arbre phylogénétique des tribus appartenant à la famille des Asteraceae. Les points noirs représentent les événements de duplication. La longueur de branche est proportionnelle au Ks (nombre de substitutions par site par an) (Barker et al. 2008)

IV.II.2. Recherche et utilisation de séquences orthologues

Il existe de nombreuses méthodes pour retrouver les séquences orthologues entre deux espèces. Ces méthodes sont classées en deux catégories : celles basées sur les arbres (tree-based methods) et celles basées sur les graphes (graph-based methods).

Les méthodes basées sur les arbres consistent à construire un arbre phylogénétique à partir d'un alignement de séquences homologues. La nature du nœud (événement de duplication ou de spéciation, **figure 1.15**) correspondant à la séquence ancestrale commune aux séquences étudiées permet de savoir si deux séquences sont orthologues ou paralogues.

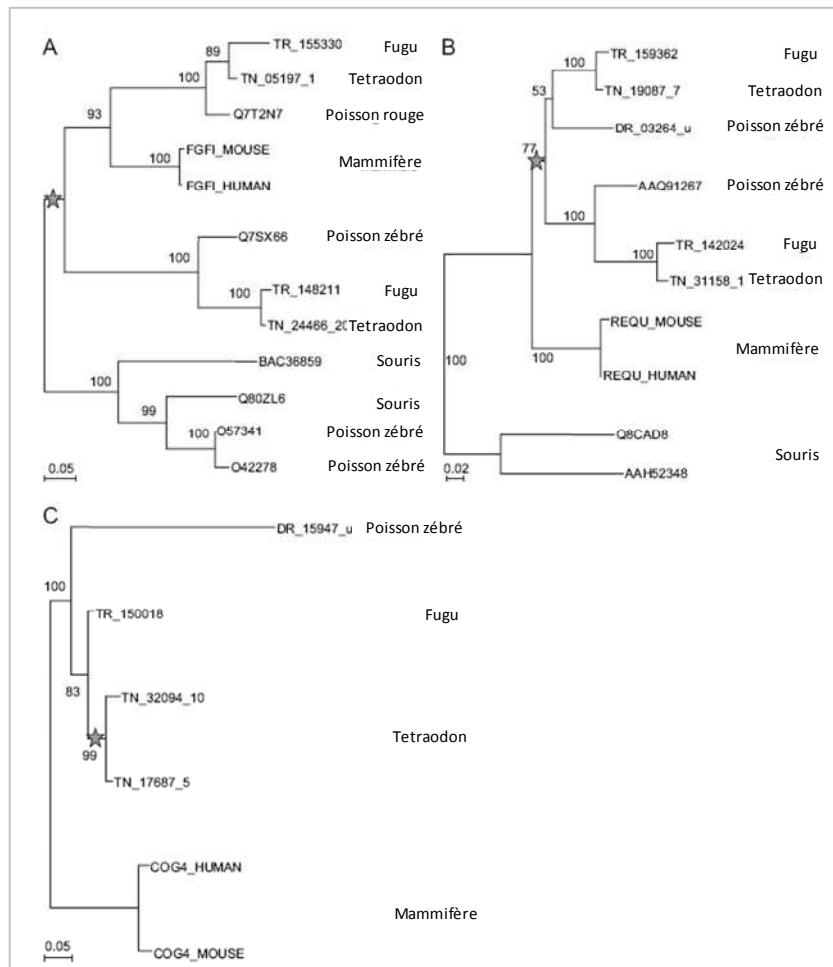


Figure 1.17 : Arbres phylogénétiques obtenus via la méthode de Neighbor-Joining. La longueur des branches est proportionnelle au nombre de substitutions par site. Chaque nœud est soutenu par un pourcentage de bootstrap. Les étoiles représentent les événements de duplications responsables de la présence de paralogues chez *Tetraodon*. Arbres construits à l'aide des homologues (A) du précurseur du facteur de croissance 18 des fibroblastes, qui met en évidence une duplication avant la séparation tétrapode- actinoptérygien, (B) d'une protéine en doigt de zinc ubi-d4 mettant en évidence une duplication chez une ancêtre des poissons et (C) d'un constituant du complexe de Golgi dont la duplication a eu lieu au sein du genre *Tetraodon* (Brunet et al. 2006).

Cette approche a été utilisée afin d'étudier les événements de duplication totale ayant eu lieu chez les ancêtres du genre *Tetraodon* (Brunet et al. 2006). Les gènes dupliqués mis en évidence ainsi que leurs homologues présents chez d'autres espèces ont été regroupés. Chaque groupe a ensuite été utilisé afin de construire un arbre phylogénétique. Suite à l'analyse de plusieurs centaines d'arbres, l'origine de trois événements de duplication totale a ainsi pu être mise en évidence. Trois exemples d'arbres ainsi obtenus sont représentés dans la **figure 1.17**.

Ce type de méthode est fortement dépendant de la qualité de l'arbre ainsi que de celle de l'alignement. De plus, les algorithmes utilisés ne sont pas toujours appropriés pour l'analyse d'une grande quantité de données.

Les méthodes basées sur les graphes reposent sur la similarité entre deux séquences calculée à partir des séquences contenues dans le génome de deux espèces. Une des méthodes les plus utilisées est le Reciprocal Best Blast Hit (RBH) dont le principe est le suivant : si deux séquences provenant de deux génomes différents se détectent réciproquement comme meilleur résultat de BLAST (Basic Local Alignment Search Tool), alors ils sont définis comme orthologues (Moreno-Hagelsieb and Latimer 2008).

De manière générale, la recherche d'orthologie peut être biaisée par la présence de xénologues, c'est-à-dire de gènes transférés de manière horizontale ainsi que par l'analyse de génome incomplet. (Kuzniar et al. 2008).

Vu la disponibilité de banques EST au sein des différentes familles botaniques, ce sont généralement ces banques qui sont utilisées pour la recherche d'orthologie. Une fois que les paires d'orthologues entre espèces ont été mises en évidence par les méthodes décrites ci-dessus, il faut positionner ceux-ci sur les génomes afin de pouvoir comparer leur position et mettre en évidence une éventuelle synténie. Cela nécessite donc de développer des marqueurs au sein de ces EST.

Le nombre de marqueurs utilisés dans les différentes études de synténie est généralement de l'ordre de quelques dizaines à une centaine (McConnell et al. 2010; Vilanova et al. 2008; Yu et al. 2004).

Par exemple, une étude de synténie au sein de la famille des Fabaceae a été réalisée à l'aide d'un total de 167 marqueurs issus de séquences EST. Afin de faire le lien entre les espèces modèles et les espèces cultivées, une carte comparative consensus a été réalisée (**figure 1.18**). Cette carte met en évidence les relations entre le génome du pois (*Pisum sativum*), du soja (*Glycine max*), du haricot (*Phaseolus vulgaris*), de la cornille (*Vigna unguiculata*), de *Medicago truncatula* (la luzerne) et de *Medicago sativa* (Choi et al. 2004).

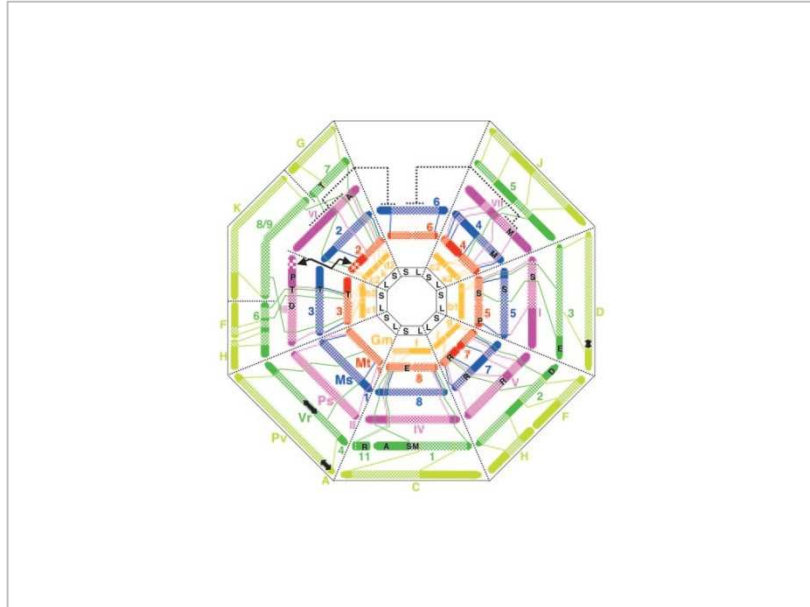


Figure 1.18: Carte comparative consensus de 6 espèces de légumineuses : *Medicago truncatula* (Mt), *Glycine max* (Gm), *Medicago sativa* (Ms), *Pisum sativum* (Ps), *Vigna radiata* (Vr) et *Phaseolus vulgaris* (Pv). S et L indiquent l'emplacement du bras long (L) et du bras court (S) des chromosomes chez *M. truncatula*. Les blocs de synténie sont dessinés à l'échelle en fonction des distances génétiques. Lignes pleines entre les génomes : réarrangements supposés. Doubles flèches : inversions supposées. S, R, T, M, P, A, D, E : Emplacement de marqueurs spécifiques de gènes qui perturbent la synténie (Choi et al. 2004)

La présence de banques EST chez les espèces étudiées, la recherche d'orthologie ainsi que le positionnement de ces séquences sur le génome sont les points clés de la cartographie comparée.

IV.III. INTÉRÊT DE LA CARTOGRAPHIE COMPARÉE

Une fois la cartographie comparée établie, celle-ci pourra être utilisée pour l'étude de l'évolution des génomes et également pour le transfert de données d'une espèce à l'autre.

IV.III.1. Etude de l'évolution des génomes

- Réarrangements chromosomiques

La cartographie comparée réalisée à l'aide de marqueurs permet de visualiser les zones conservées ainsi que les zones dissemblables entre deux espèces. Ces différences sont dues aux réarrangements à grande échelle de type délétion, duplication, inversion, translocation, fusion et fission qui peuvent survenir au sein des chromosomes. Les mécanismes à l'origine de ces événements sont la cassure et la réunion des chromosomes. Certains changements dans la structure des chromosomes résultent souvent d'une cassure chromosomique et les extrémités cassées du chromosome sont hautement réactives. En effet, elles présentent une forte tendance à s'unir à d'autres extrémités cassées.

- Evolution des génomes chez les angiospermes

Les études d'orthologues et de paralogues au sein des différentes familles d'angiospermes ont permis de reconstruire les génomes ancestraux et de retracer l'histoire évolutive de ces différentes familles (**figure 1.19**) (Salse 2012).

Plusieurs événements de polyploïdisation ont ainsi pu être mis en évidence. Ces événements sont généralement suivis d'un grand nombre de réarrangements chromosomiques (fusions et fissions) qui permettent de réduire la taille du génome et le nombre de chromosomes. Les régions où ont eu lieu ces réarrangements sont dans les génomes actuels des régions fragiles où l'on peut observer un taux de recombinaison anormal dû aux nombreuses mutations occasionnées par les réparations. D'un point de vue évolutif, ces régions sont des réservoirs de nouveautés engendrant ainsi de nouveaux phénotypes (Salse 2012).

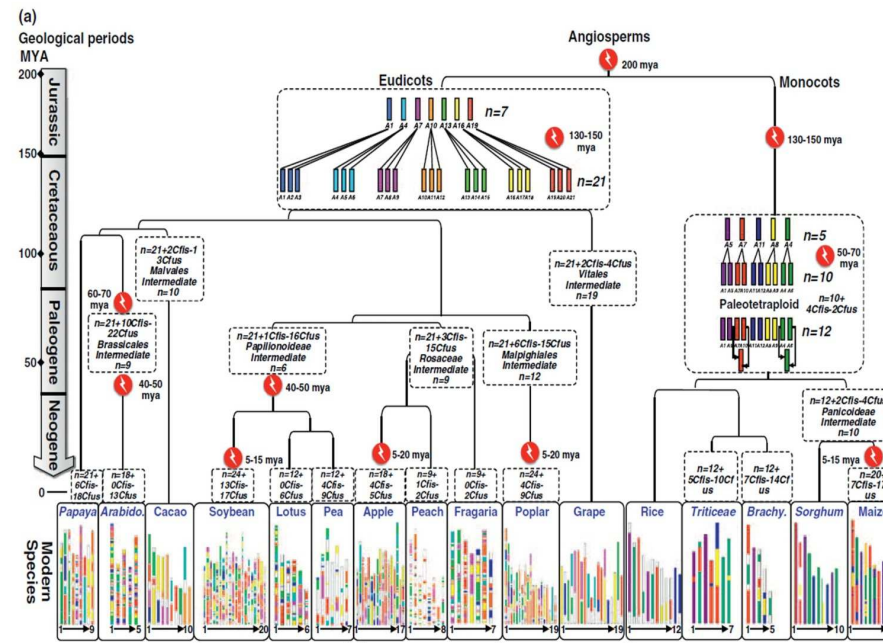


Figure 1.19: Représentation des génomes ancestraux, intermédiaires et actuels de différentes espèces d'angiospermes. 🚩 : Événements de polyploïdisation. L'équation caryotypique permettant d'expliquer le nombre de chromosomes de chaque génome est indiquée. Cfis : fission chromosomique. Cfus : fusion chromosomique. (Salse 2012)

- Exemple de cartographie comparé pour l'étude de l'évolution des génomes

Une étude de la famille des Poaceae a permis d'identifier les différents réarrangements chromosomiques responsables de l'organisation des génomes actuels. La famille des Poaceae regroupe les cinq espèces de céréales les plus importantes d'un point de vue socio-économique. Celles-ci sont regroupées dans trois sous-familles : Les Panicoideae (le sorgho et le maïs), les Ehrhartoideae (le riz) et les Pooideae (le blé et l'orge) (**figure 1.20**).

Afin d'étudier les relations entre le génome de ces différents espèces, une cartographie comparée a été réalisée sur base de marqueurs RFLP (Moore et al. 1995). Suite à l'important développement des données génétiques pour cette famille, cette étude a été complétée par l'analyse du génome séquencé du riz et du sorgho, par la carte physique du maïs ainsi que par les nombreuses EST cartographiées pour le genre Triticum (Bolot et al. 2009a).

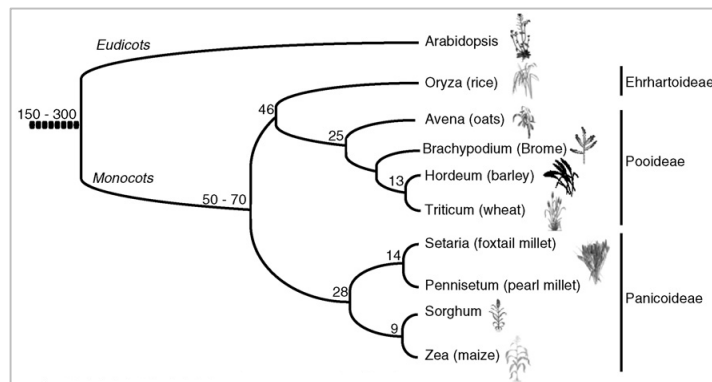


Figure 1.20 : Arbre phylogénétique de la famille des Poaceae. Le temps séparant deux espèces d'un ancêtre commun est indiqué sur les branches en millions d'années (Bolot et al. 2009)

L'étude de la macrolinéarité entre ces espèces a montré que, malgré le nombre variable de polyploïdies et de chromosomes, l'ordre des marqueurs et des séquences semble hautement conservé. L'alignement des différents génomes a été représenté sous forme de « crop circles » permettant de mettre en évidence les relations entre chromosomes orthologues chez les cinq espèces majeures (**figure 1.21**).

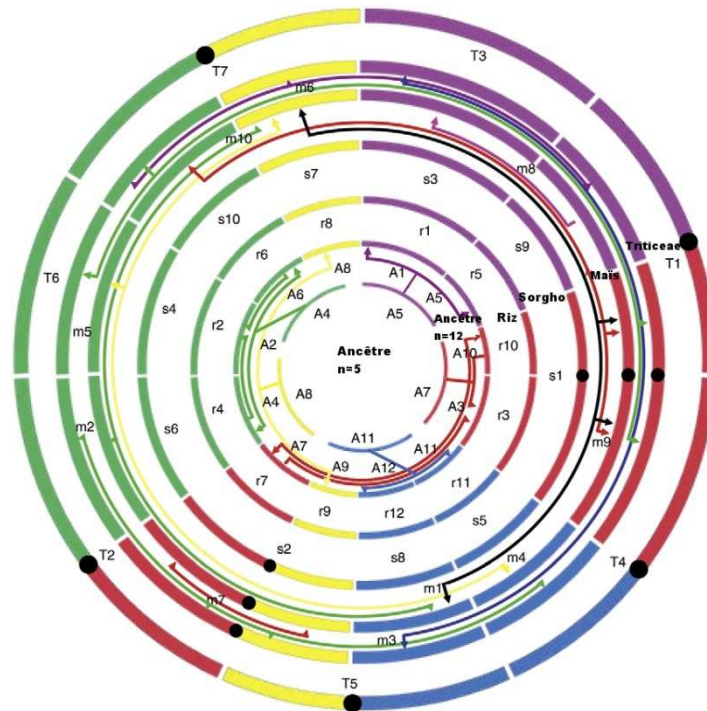


Figure 1.21 : « Crop circles » représentant l'alignement des génomes d'espèces appartenant à la famille des Poaceae. Les chromosomes des Triticeae, du maïs, du sorgho et du riz sont représentés sous forme de cercles en fonction de la taille de leur génome. Le génome du riz se situe le plus au centre car il est le plus proche des génomes ancestraux. Le génome du maïs est représenté par les 20 chromosomes originaux obtenus après l'événement de tetrapolyploïdisation. Les 17 translocations qui ont suivi cet événement donnant lieu aux 10 chromosomes réarrangés sont représentées par les flèches de couleur. La fusion des chromosomes est représentée par les ronds noirs. Les deux cercles internes représentent le génome ancestral des graminées ($n=5$) et le génome de l'ancêtre intermédiaire ($n=12$). Les 5 couleurs des chromosomes font référence aux 5 chromosomes ancestraux et les flèches colorées représentent le lien entre le génome de l'ancêtre $n=5$ et le génome de l'ancêtre $n=12$ obtenu après un événement de duplication totale (Bolot et al. 2009)

Le cercle interne représente l'ancêtre commun présumé. Il est composé de cinq chromosomes. Le deuxième cercle représente un ancêtre intermédiaire qui aurait subi, il y a 50-70 MYA, une duplication totale du génome (WGD, whole genome duplication), suivie de deux translocations chromosomiques amenant le nombre de chromosomes à 12. Le troisième cercle représente le génome du riz ($n=12$), très proche de l'ancêtre intermédiaire.

L'ancêtre de la sous-famille des Panicoideae aurait subi deux fusions chromosomiques (rond noir) amenant le nombre de chromosomes à 10. Le génome du sorgho ne semble pas avoir été beaucoup modifié par rapport à cet ancêtre. L'ancêtre du maïs, lui, aurait subi une duplication totale du génome donnant un intermédiaire possédant 20 chromosomes (représenté par le double cercle). Suite aux 17 fusions chromosomiques (représentées par les flèches entre le double cercle), le nombre de chromosomes est redescendu à 10 comme observé chez le maïs actuellement.

Concernant le genre *Triticum*, cinq fusions chromosomiques à partir de l'ancêtre intermédiaire (n=12) ont abouti au génome actuel du blé et de l'orge avec un nombre de base de chromosomes égal à 7.

IV.III.2. Transfert d'informations génétiques

Comme décrit ci-dessus, la famille des Poaceae présente une forte macrosynténie entre espèces. Cette spécificité a été utilisée pour la caractérisation de QTL chez le blé. Un QTL lié à la viscosité des pentosans, un critère influençant la qualité de la farine, a été identifié dans une région recouvrant 80 cM du chromosome 7a et co-localisé avec douze marqueurs associés à des EST. L'orthologue de ces EST a été identifié sur le génome du riz pour lequel les ressources génétiques sont plus nombreuses. Deux régions chromosomiques ont ainsi été délimitées dans lesquelles se trouvent 124 séquences. Ces séquences ont été annotées et sont des gènes candidats potentiels pour la caractérisations du QTL (Ouraishi et al. 2009).

Une autre étude du même genre a été réalisée chez la cornille, une espèce de la famille des Fabaceae. Plusieurs gènes candidats pouvant expliquer la variabilité au sein de QTL associés à la résistance à un pathogène fongique ont été identifiés grâce à l'analyse de la zone de synténie chez deux espèces modèles : *M. truncatula* et *G. max* (Muchero et al. 2011).

Comme ces exemples le montrent, un des avantages majeurs de la cartographie comparée est donc le fait qu'une fois les zones de synténie identifiées entre deux génomes, il est possible de transférer les données génétiques d'une espèce à l'autre. Dans le cas d'espèces possédant des ressources génétiques limitées comme la chicorée, cela permet d'ouvrir de nouvelles perspectives de recherche.

Objectifs

L'objectif général de ce travail est de construire des outils permettant d'améliorer la sélection génétique de la chicorée via trois aspects : le développement de marqueurs moléculaires, l'étude de l'organisation du génome et le développement des données génétiques disponibles pour la chicorée industrielle.

A ce jour, peu de marqueurs sont disponibles pour la chicorée industrielle. Nous avons donc entrepris de développer des marqueurs spécifiques de locus, co-dominants et facilement utilisables. Afin de réaliser cet objectif, les séquences EST disponibles pour la chicorée ont été analysées pour mettre au point des marqueurs de type SSR et SNP.

Afin de mieux comprendre l'organisation d'un génome, une carte génétique indique la position de gènes les uns par rapport aux autres, permet d'identifier des QTL ou encore de réaliser des introgressions pour l'amélioration des variétés. Trois cartes génétiques ont été publiées pour le genre *Chicorium*. Malheureusement, les données de ces différentes cartes n'ont pas pu être utilisées chez les individus à analyser dans ce travail. Une nouvelle carte génétique a donc été construite en collaboration avec l'UCL et la société Cosucra à l'aide d'une population F2 nommée SAMCHIC de 247 individus ainsi que de marqueurs de type AFLP, SSR et SNP.

Le troisième objectif de ce travail était de développer les données génétiques disponibles pour la chicorée industrielle. L'idée était de transférer des données génétiques provenant d'une espèce mieux étudiée et proche de la chicorée. La laitue répond à ces deux critères puisque le genre *Lactuca* est phylogénétiquement proche de *Chicorium* et à ce jour, de nombreuses données telles que les cartes génétiques et physiques, les banques EST ou encore des analyses QTL sont disponibles au sein de ce genre. Afin de savoir si ces données pouvaient être transférées chez la chicorée, une cartographie comparée a été réalisée entre les deux espèces pour mettre en évidence une éventuelle synténie.

La stratégie utilisée pour réaliser ces trois objectifs est la suivante. Dans un premier temps, afin de mettre en évidence des régions de type SSR et SNP, les banques EST de chicorée industrielle ainsi que des séquences potentiellement orthologues avec des EST de laitue ont été analysées.

A partir de ces analyses, des marqueurs géniques et co-dominants ont été mis au point. Ces marqueurs sont venus compléter un jeu de marqueurs AFLP développé en collaboration avec l'UCL et un jeu de marqueurs SSR génomiques développé par Cosucra. Cet ensemble de marqueurs a donc été utilisé pour génotyper la population de ségrégation SAMCHIC et pour construire la carte génétique de la chicorée.

Enfin, les marqueurs géniques positionnés sur la carte ont été utilisés pour la recherche d'orthologues potentiels avec les EST positionnées sur la carte physique et les cartes génétiques de la laitue. Une carte comparée a ainsi pu être construite et l'analyse de la synténie entre les deux espèces a pu être initiée. La **figure 1.22** résume cette stratégie et sa répartition en différents chapitres.

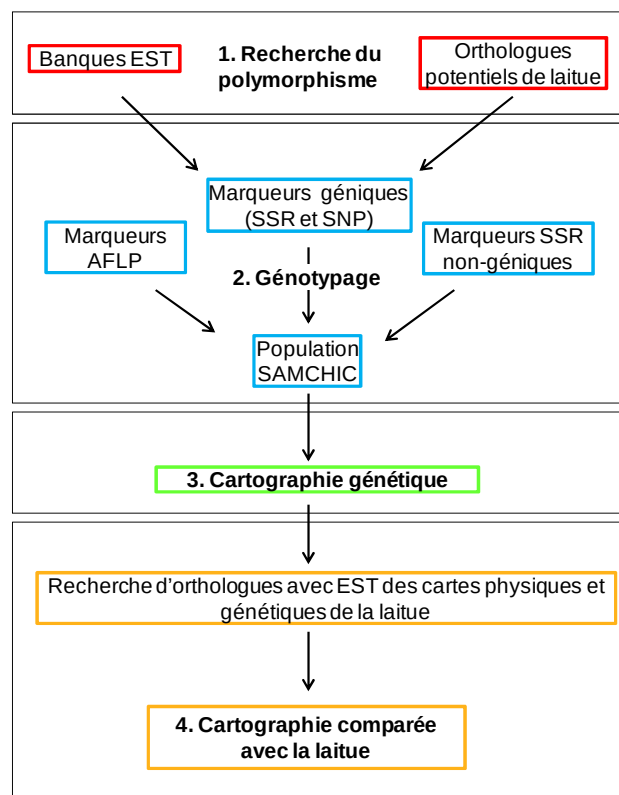


Figure 1.22 : Vue d'ensemble de la stratégie utilisée. Les quatre encadrés noirs représentent les quatre chapitres.

Chapitre 2

Recherche de polymorphisme moléculaire

Chapitre 2. MISE EN ÉVIDENCE DU POLYMORPHISME MOLÉCULAIRE

I. BUT DU CHAPITRE

Afin de développer un grand nombre de marqueurs moléculaires, les banques EST de chicorée ont été utilisées pour mettre en évidence du polymorphisme de type SSR et SNP. L'utilisation d'outils bioinformatiques a permis de réaliser cette étape à moindre coût et en un minimum de temps. Cette démarche a permis au final de mettre en évidence un grand nombre de sites SSR et SNP potentiellement polymorphes chez la chicorée. Dans le présent chapitre, les résultats de la recherche de motifs SSR et SNP au sein des banques EST de chicorée via les outils bioinformatiques sont détaillés.

Ensuite, afin de réaliser la cartographie comparée entre la chicorée et la laitue, une recherche d'orthologues a été réalisée entre les banques EST des deux espèces. Les séquences sélectionnées ont été analysées et séquencées chez la chicorée afin de mettre en évidence du polymorphisme utilisable pour la cartographie. Les différentes étapes de cette démarche sont également décrites dans ce chapitre.

II. POLYMORPHISME AU SEIN DES BANQUES EST

II.1. LES BANQUES EST CHEZ LA CHICORÉE

Les banques EST, comme expliqué dans le point II.IV du chapitre 1, sont une source de données génétiques riches en polymorphisme. En effet, ces séquences étant issues d'ARN_m provenant généralement de plusieurs individus eux-mêmes potentiellement hétérozygotes, chaque gène peut se retrouver en plusieurs copies dans la banque. L'alignement de plusieurs copies d'une même séquence permet de mettre en évidence le polymorphisme dans la banque de données.

A ce jour, trois banques EST de *C. intybus* sont disponibles :

- Banque EST Phytomol: 12.226 EST
- Banque EST de Davis (notée « Banque EST D ») : 38.323 EST
- Banque EST de Lille (notée « Banque EST L ») : 3.348 EST

La banque EST Phytomol (Dauchot et al. 2009) est issue de 10 librairies d'ADNc provenant de racines de chicorée industrielle (*C. intybus* var. *sativum* cultivar « Arancha ») prélevées tout au long de la saison 2002-2003. Deux autres librairies provenant de nodules et de feuilles ont été ajoutées. Au final, 12.226 EST ont été séquencées. Les analyses bioinformatiques réalisées sur cette banque seront détaillées ultérieurement.

La banque EST de Davis a été réalisée dans le cadre du Compositae Genome Project (<http://compgenomics.ucdavis.edu/>). Ce projet, mis en place par l'Université de Californie à Davis aux USA, a pour but de développer un catalogue de séquences génétiques ainsi que des cartes génétiques détaillées pour les espèces appartenant à la famille des Asteraceae (ou Compositae). Cette banque comporte 38.323 EST issues de racines, de feuilles et de fleurs de chicorée (*C. intybus* Ames 3210#2, http://www.ars-grin.gov/npgs/acc/acc_queries.html).

L'université de Lille, en France, a également développé une banque EST à partir de feuilles de deux plants de chicorée obtenus par embryogénèse somatique (Legrand et al. 2007). Cette banque se compose de 3.348 EST.

Au final, un jeu de 53.897 EST est disponible pour *C. intybus*. Ces séquences vont être analysées afin de mettre en évidence du polymorphisme microsatellite et SNP.

II.II. MISE EN ÉVIDENCE DE MARQUEURS MICROSATELLITES

II.II.1. Introduction

Le but de cette partie du travail était de mettre en évidence et de sélectionner des séquences possédant des motifs SSR au sein d'une banque d'EST dans l'optique de développer des marqueurs pour la cartographie.

La première étape consistait à isoler les séquences possédant les motifs SSR à l'aide d'outils bioinformatiques. Ensuite, ces séquences ont été annotées afin de sélectionner les SSR localisés dans des gènes d'intérêt. Les microsatellites ont également été sélectionnés suivant le type de motif et la taille de la séquence répétée afin d'augmenter les chances d'avoir des marqueurs polymorphes. Parmi cette sélection, seules les séquences permettant la conception d'amorces dans les régions flanquantes du SSR ont été gardées.

Cette stratégie a donc permis d'obtenir un jeu de séquences possédant des motifs SSR associés à une fonction d'intérêt dans laquelle des amorces ont été définies pour le génotypage des loci microsatellites.

II.II.2. Matériel et méthodes

II.II.2.1. Création d'une banque d'unigènes

L'assemblage des trois banques EST décrites ci-dessus a été réalisé à l'aide du programme EGassembler (<http://egassembler.hgc.jp/>) (Masoudi-Nejad et al. 2006). L'option « Enable Repeat Masking Process » a été désélectionnée afin de ne pas réduire la taille des motifs répétés.

II.II.2.2. Recherche de motifs microsatellites

Le programme Sputnik (<http://services.cbib.u-bordeaux2.fr/pise/sputnik.html>) (Abajian 1994) a été utilisé pour mettre en évidence les motifs SSR au sein des unigènes. La taille minimale de la séquence répétée était de deux nucléotides et la séquence totale du microsatellite était de minimum 12 paires de bases. Seuls les microsatellites parfaits ont été retenus.

II.II.2.3. Annotation des EST

Les unigènes ont été blastés et annotés à l'aide du programme Blas2Go (<http://www.blast2go.com/b2ghome>) (Conesa et al. 2005). Les critères suivants ont été sélectionnés : Database : nr. Hits : 20. E value : $1e^{-3}$. Blastx.

II.II.2.4. Conception des amorces flanquant les SSR

Les amorces servant à amplifier les régions microsatellites ont été définies à l'aide des programmes Webtroll (<http://wsmartins.net/websat/>) et Primer3 (<http://frodo.wi.mit.edu/>) (Rozen and Skaletsky 2000) avec les critères suivants :

- Tm (température de fusion) entre 50 et 60°C
- Taille des amorces : ± 20 paires de bases
- Pourcentage en GC entre 45 et 55

Chaque paire d'amorces a été analysée via le programme Vector NTI afin de vérifier qu'il n'y avait pas apparition de structures secondaires avec un ΔG inférieur à -5 kcal/mol.

II.II.3. Résultats et discussion

II.II.3.1. Création d'un jeu d'unigènes

Vu le taux élevé de mutations au sein des SSR (entre 10^{-2} et 10^{-6} par locus et par génération) et même si celui-ci est variable suivant la nature, la taille et la localisation du microsatellite, on s'attendait à observer du polymorphisme pour une majorité des marqueurs SSR. La mise en évidence de motifs microsatellites au sein de séquences EST ne nécessitait donc pas l'alignement de différents allèles. Il était donc

plus facile de travailler avec un jeu d'unigènes. Une banque d'unigènes est une base de données reprenant les séquences consensus obtenues après alignement d'EST similaires ainsi que les EST présents en un seul exemplaire dans la banque de données initiale c'est-à-dire, les singletons.

Le programme EGassembler (Masoudi-Nejad et al. 2006) a été utilisé pour générer ce jeu d'unigènes à partir des séquences d'EST provenant des trois banques décrites ci-dessus. Ce programme est divisé en cinq étapes. La première étape est une étape de nettoyage qui comporte la suppression des poly A et T, des séquences bordantes de mauvaise qualité, des séquences trop courtes ou encore de faible complexité. Les trois étapes suivantes consistent à masquer au sein des séquences restantes les motifs répétés, les vecteurs ainsi que les séquences provenant d'organelles. Enfin, la dernière étape est réalisée par le programme CAP3 (Huang and Madan 1999) qui permet de détecter les chevauchements entre les séquences ainsi que leur qualité. Sur base de ces analyses, des contigs regroupant des séquences similaires sont formés et alignés générant ainsi pour chaque groupe, une séquence consensus. Les paramètres par défaut pour la création de contigs sont un chevauchement de minimum 65 nucléotides et 80 % d'identité. Au final, on obtient un ensemble de séquences consensus issus des contigs et un ensemble de singletons ; les deux fichiers sont rassemblés afin d'obtenir le jeu d'unigènes.

Par défaut, EGassembler réduit les motifs répétés à la plus petite version présente dans le contig. Cela pouvait poser un problème dans notre sélection. En effet, les motifs SSR ont été sectionnés suivant leur taille car celle-ci est liée au polymorphisme. Le fait de réduire les SSR à leur plus petite forme risquait donc de biaiser nos résultats. Le jeu d'unigènes a donc été créé sans réduction de la taille des séquences répétées (disable masking process).

Au final, un jeu de 22.038 unigènes a été obtenu. C'est sur base de ce jeu de données que les motifs microsatellites ont été recherchés.

II.II.3.2. Recherche de motifs microsatellites

Afin de sélectionner les séquences possédant des motifs SSR, nous avons réalisé une analyse bioinformatique à partir de la banque d'unigènes.

- Stratégie

Le programme Sputnik permet, à partir d'un fichier en format « fasta » pouvant contenir plusieurs milliers de séquences, d'isoler celles qui présentent des séquences répétées de type microsatellite. Ce programme permet de sélectionner plusieurs critères de recherche tels que les tailles minimale et maximale du motif répété, la taille totale du SSR ainsi que le pourcentage d'imperfections dans le microsatellite. Les critères sélectionnés pour notre analyse sont décrits et expliqués ci-dessous.

La taille minimale du motif répété sélectionné était de deux nucléotides. En effet, la discrimination des différents allèles d'une séquence comprenant des répétitions mononucleotidiques n'est pas évidente. Lors de la PCR, la Taq polymérase peut ajouter une adénine en fin de séquence. Il est donc difficile de faire la différence entre deux allèles de taille différente et deux allèles identiques avec ou sans adénine supplémentaire.

La taille totale du SSR influence directement le taux de polymorphisme rencontré dans une population. En effet, comme décrit dans le point II.II.2 du chapitre 1, le taux de mutations au sein d'un microsatellite augmente avec la taille de celui-ci. Cela est dû au fait que le nombre de « slippages » est plus important en présence d'un plus grand nombre de répétitions. Afin d'augmenter les chances que les microsatellites sélectionnés soient polymorphes, la taille minimale de la séquence répétée était d'au moins 10 nucléotides : soit minimum cinq répétitions pour un dinucléotide et quatre pour un trinucleotide.

Seuls les microsatellites parfaits, c'est-à-dire composés uniquement de répétitions identiques et non entrecoupés par des mutations ponctuelles, ont été retenus. Les SSR imparfaits possèdent généralement une taille plus petite puisqu'ils sont issus de SSR parfaits qui ont été coupés par des mutations ponctuelles. Par ce fait, ils ont souvent des taux de mutations plus faibles, ce qui rend ces marqueurs potentiellement moins polymorphes. De plus, les SSR imparfaits sont

plus difficiles à analyser : la présence de mutations ponctuelles biaise le calcul du nombre de répétitions et donc l'identification des différents allèles.

- Analyse des résultats

Au total, 3.443 séquences contenant un microsatellite ont été sélectionnées. Cela représente 15,6% de notre banque d'unigènes. Ces résultats sont comparables aux données obtenues par N. Dauchot (Dauchot 2007) lors de l'analyse de la banque EST Phytomol. En effet, 15% du jeu d'unigènes généré à partir des 12.226 EST comportaient des motifs microsatellites.

Parmi ces séquences sélectionnées, 59% des SSR étaient de type trinuécléotidique (**figure 2.1**). Les mutations au sein de ce type de SSR n'engendrent pas de décalage dans la phase de lecture. Ces mutations subissent donc moins de pression de sélection, ce qui explique que ce type de SSR se retrouve en plus grand nombre dans les régions transcrites.

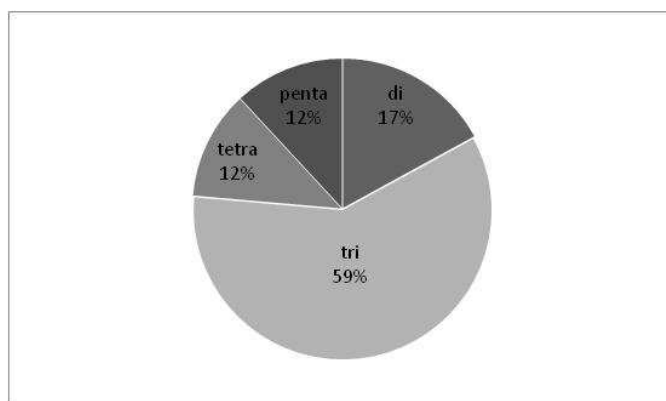


Figure 2.1 : Pourcentage de microsatellites de type di-, tri-, tetra- et pentanucléotidique au sein de 3.443 EST provenant d'un jeu de 22.038 unigènes de chicorée

La **figure 2.2** représente le nombre de séquences possédant des motifs SSR en fonction du nombre de répétitions observées. La présence de motifs trinuécléotidiques de quatre répétitions était largement supérieure aux autres motifs. Concernant les SSR de type dinuécléotidique (en rouge), la majorité de ces séquences possédait six répétitions, soit une taille totale de 12 pb. Pour les penta- et tétranuécléotides (en mauve et bleu), la plupart des séquences

possédait trois répétitions, soit une taille totale de respectivement 12 et 15 pb. La majorité des SSR, tous types confondus, possède donc une taille de 12 pb. Cette taille pourrait correspondre à une taille optimale pour laquelle le rapport insertion/délétion lors du « slippage replication » est atteint.

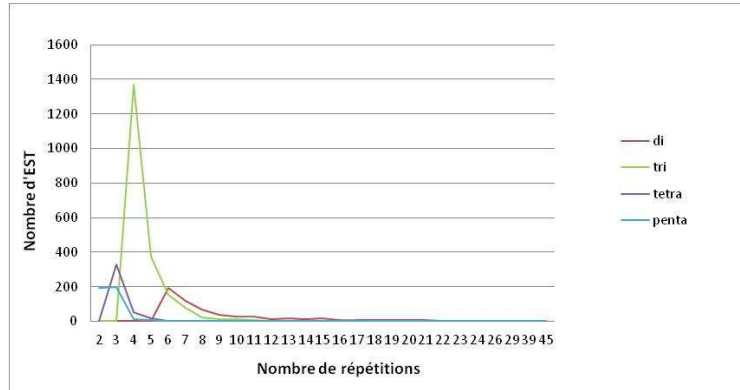


Figure 2.2 : Nombre de séquences EST en fonction du nombre de répétitions contenues dans les microsatellites di-, tri-, tetra- et pentanucléotidiques

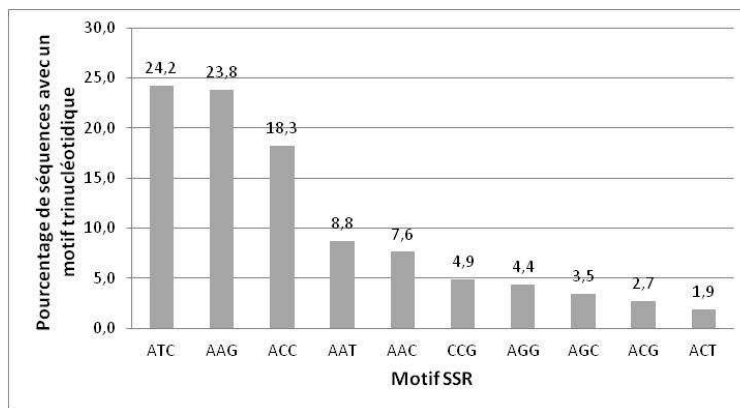


Figure 2.3 : Répartition des motifs SSR au sein des séquences possédant des microsatellites trinucleotidiques

Concernant la nature des motifs, 48% des SSR trinucleotidiques étaient composés de motifs ATC et AAG (**figure 2.3**), ce qui représentait 30% des séquences totales (**figure 2.4**). Cette même distribution a été observée préalablement pour l'analyse de la banque Phytomol (chapitre 1, **figure 1.5**)(Dauchot 2007).

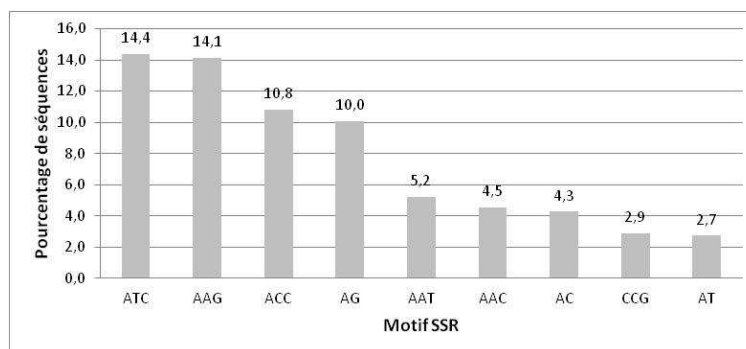


Figure 2.4 : Répartitions des motifs SSR les plus fréquents au sein des 3.443 unigènes de chicorée.

Pour les motifs dinucléotidiques (**figure 2.5**), les données observées étaient différentes par rapport aux données obtenues suite à l'analyse de la banque Phytomol (chapitre 1, **figure 1.5**). Dans notre cas, 59% des SSR dinucléotidiques possédaient des motifs AG et 16% des motifs AT contre respectivement 33% et 41% dans la banque Phytomol. Cette différence peut s'expliquer soit par le fait que les critères de sélection étaient légèrement différents entre les deux analyses (dans notre cas, pas de SSR imparfait) ou encore par le fait que la provenance des tissus utilisés pour générer les banques EST était différente et donc que les séquences transcrites étaient également distinctes, ce qui a engendré un biais dans le type de motif microsatellite observé.

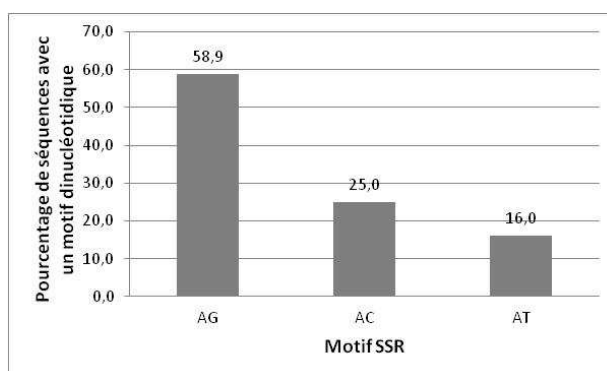


Figure 2.5 : Répartition des motifs SSR au sein des séquences possédant des microsatellites dinucléotidiques

Afin d'approfondir cette réflexion, la recherche de SSR a été réalisée séparément pour les trois banques EST c'est-à-dire, la banque Phytomol, la banque de Lille (banque L) et la banque de Davis (Banque D) à l'aide de critères de sélection identiques (taille minimale du SSR : 10 pb et uniquement des SSR parfaits) (**figure 2.6**). Un jeu d'unigènes a été généré pour chacune des banques et analysé via le programme Sputnik. Les motifs AC, AG et AT représentent chacun environ un tiers des motifs dinucléotidiques de la banque Phytomol. Une répartition beaucoup moins homogène a été observée pour les banques EST L et D où le nombre de motifs AG est beaucoup plus important et où le nombre de motif AT est beaucoup plus faible.

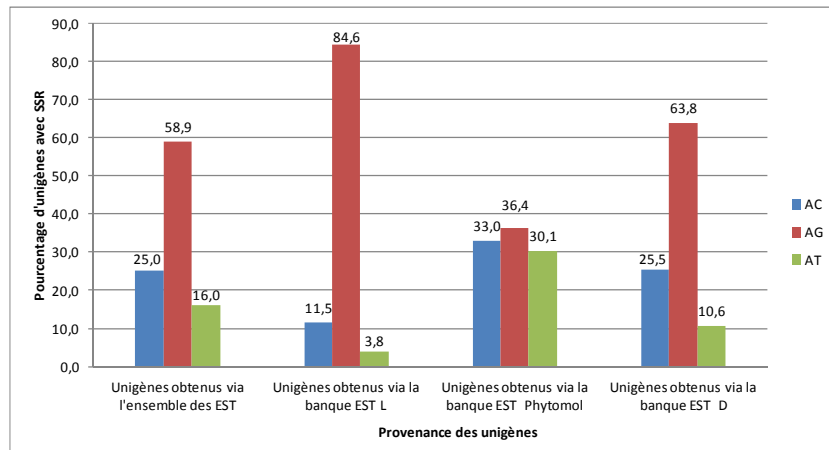


Figure 2.6: Pourcentage d'unigènes de chicorée possédant des SSR dinucléotidiques de type AC, AG et AT. L'analyse a été réalisée à partir de l'ensemble des EST disponibles pour la chicorée (22.038 unigènes), de la banque EST L (2.103 unigènes), de la banque Phytomol (8.116 unigènes) et de la banque EST D (21.309 unigènes).

La différence de distribution des motifs dinucléotidiques observée dans les différentes banques ne peut pas s'expliquer par une différence entre les critères de sélection des motifs SSR puisque ceux-ci sont identiques (**figure 2.6**). Plusieurs hypothèses sont envisagées afin d'expliquer cette divergence.

Tout d'abord, les critères de nettoyage (suppression des poly A et T, des régions bordantes de mauvaise qualité, des séquences trop courtes ou encore de faible complexité) des séquences brutes peuvent diverger d'une banque à l'autre. Lors de cette étape les répétitions AT aux extrémités sont généralement supprimées. Il se peut que les

critères de suppression de ces régions aient été plus contraignants dans les banques L et D que dans la banque Phytomol. Cela pourrait expliquer le pourcentage plus élevé de motifs AT dans la banque Phytomol.

Une deuxième hypothèse est que, suivant la Taq polymérase utilisée, la présence de motifs répétés AT (pour les banque EST L et D) ou AG (pour la banque Phytomol) génère des séquences de mauvaise qualité qui ont alors été supprimées.

Enfin, comme émis plus haut, la différence de distribution entre les banques EST pourrait s'expliquer par une différence au niveau de la nature des séquences transcrites entre les trois banques. En effet, si les séquences sont différentes, il est possible que les SSR contenus dans celle-ci aient des motifs différents. Cette différence de séquence peut s'expliquer par la nature des tissus utilisés. La banque Phytomol a été créée à partir majoritairement de racines (10 banques sur 12) mais également de nodules (1/12) et de feuilles (1/12). La banque L a été construite à partir de feuilles provenant de deux plants issus d'embryogénèse somatique. Enfin, la banque D est issue de séquences provenant de racines, de fleurs et de feuilles.

Les motifs dinucléotidiques GC ne sont pas observés au sein des SSR mis en évidence dans notre analyse. Ce motif est très peu ou pas présent dans les EST de la laitue (Riar et al. 2011; Simko 2009), du génome de l'humain, de la drosophile, d'Arabidopsis (Li et al. 2004) ainsi qu'au sein des EST d'autres plantes modèles (Victoria et al. 2011). Ce phénomène pourrait s'expliquer par la méthylation des cytosines présentes dans ces régions. Cette modification augmente les chances de mutation de la cytosine en thymine via une déamination (Garrett and Grisham 2000; Katti et al. 2001). Ce mécanisme engendrerait donc des SSR trop imparfaits pour être retenus lors des analyses

II.II.3.3. Sélection des séquences suivant leur fonction

- Intérêt de la démarche

Les séquences à partir desquelles sont développés les marqueurs moléculaires peuvent être sélectionnées suivant leur fonction (gènes candidats fonctionnels) ou suivant leur position stratégique sur le génome (gènes candidats positionnels). Lors de cette étape, les EST ont

été sélectionnées suivant leur fonction. En effet, le développement de marqueurs moléculaires au sein d'EST permet d'obtenir des marqueurs se trouvant au sein de la séquence transcrite des gènes. Ces séquences peuvent être associées à une fonction potentielle via la recherche d'homologues dont la fonction est connue.

Comme expliqué dans le point II.II.3 du chapitre 1, les mutations au sein des SSR présents dans les régions transcrites d'un gène peuvent avoir un impact direct ou indirect sur la séquence de la protéine codée et donc sur sa structure et sur son activité. Les marqueurs microsatellites peuvent donc être associés au phénotype engendré par ces mutations. D'autre part, sans en être directement responsable, un marqueur microsatellite peut être lié physiquement à des mutations engendrant des variations phénotypiques. Dans ce cas, suivre ce marqueur revient à suivre le phénotype.

Sur une carte génétique, la co-localisation de ce type de marqueur avec des QTL permet d'obtenir des gènes candidats positionnels potentiellement responsables ou simplement liés aux variations de ce trait quantitatif. Cette approche a été utilisée dans différentes études en vue de caractériser des QTL. Par exemple, chez l'orge, un marqueur EST-SSR a été localisé à proximité d'un locus lié à la stérilité mâle (*msg6*) ainsi qu'à proximité d'un QTL lié au poids de l'épi. Ce marqueur semble être un bon outil en sélection afin de déterminer les individus porteurs des allèles d'intérêt (Emebiri 2010). Une autre étude du même type a été réalisée chez la palourde *Meretrix meretrix*. Trois marqueurs de type EST-SSR sont associés à un QTL lié à la taille du coquillage. Ces marqueurs montrent des différences de fréquence allélique entre les populations étudiées. De plus, leur fonction suggère qu'ils pourraient être impliqués dans le processus de croissance de cet organisme (Lu et al. 2012).

Afin de développer des EST-SSR dont la fonction a un intérêt particulier chez la chicorée, nous nous sommes focalisés sur des fonctions ayant un lien avec les critères de sélection, c'est-à-dire le métabolisme des sucres, la résistance aux pathogènes et la morphologie racinaire.

- Principe de l'annotation des EST possédant des motifs SSR

Le programme Blas2Go permet d'annoter une fonction à un grand nombre de séquences. La première étape consiste à rechercher les séquences homologues via la fonction « BLAST » au sein des banques de données publiques telles que NCBI (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>).

Ensuite, lors du « mapping », chaque séquence homologue est associée à différentes ontologies de gènes (GO, Genes Ontology) qui la définissent. Les GO ont été créées afin de structurer la description des gènes dans le cadre d'une ontologie commune à toutes les espèces (TheGeneOntologyConsortium 2008). Trois catégories principales de GO peuvent être définies.

- la fonction moléculaire (F) associée au gène
- le processus biologique (P) dans lequel il est impliqué
- le compartiment cellulaire (C) dans lequel se trouve la protéine codée par le gène.

La dernière étape dite d'annotation regroupe les différentes GO obtenues et ne garde que celles qui sont les plus représentatives du groupe de séquences homologues. Chaque séquence analysée possède au final un certain nombre d'ontologies qui la définissent.

- EST sélectionnées

Au total, plus de 1.400 séquences soit environ 40% des séquences avec motifs SSR ont été sélectionnées pour leurs annotations fonctionnelles.

Concernant la production de sucre, les processus biologiques sélectionnés étaient l'allocation du carbone, le transport et la photosynthèse. Pour la résistance aux maladies, les gènes impliqués dans la réponse au stress ainsi que dans la structure de la paroi ont été choisis. Les séquences jouant un rôle dans la morphologie racinaire ont également été prises. Enfin, de manière plus large, les facteurs de transcription, les protéines impliquées dans les modifications post-traductionnelles ainsi que celles impliquées dans la transduction du signal ont été ajoutées à notre sélection.

Au final, cette sélection a permis de se focaliser sur des gènes ayant plus de chances d'être impliqués directement ou indirectement dans

les mécanismes en relation avec des phénotypes d'intérêt pour la chicorée industrielle.

II.II.3.4. Sélection des séquences suivant les caractéristiques intrinsèques des microsatellites

En plus de la sélection des SSR dans le programme Sputnik, d'autres critères ont été ajoutés afin de sélectionner les séquences ayant le plus de chances de donner des marqueurs microsatellites valables et polymorphes.

Seuls les microsatellites de type di- et trinuécléotique ont été retenus. En effet, les SSR de type trinuécléotidique présentent plus de polymorphisme au sein des régions codantes que les autres motifs puisque l'ajout ou la suppression d'une répétition ne change pas le cadre de lecture de la protéine. Ils ne sont pas contre-sélectionnés et ce sont donc les plus abondants. Les microsatellites de type dinuécléotidique ont également été sélectionnés car ce sont les deuxièmes motifs les plus nombreux au sein des séquences codantes (voir chapitre 1 point II.II.2.5).

Enfin, sur chaque séquence sélectionnée, il était nécessaire de pouvoir définir des amorces bordant le microsatellite afin de réaliser le génotypage des individus. Cette étape était très restrictive. En effet, bon nombre de séquences étaient inutilisables car soit les régions bordantes étaient trop courtes, soit trop riches ou trop pauvres en GC, soit les amorces définies formaient des structures secondaires.

Au final, à partir des 1.400 séquences, 109 séquences répondant aux critères décrits ci-dessus ont été sélectionnées pour le développement de marqueurs microsatellites. Ces marqueurs ont été testés et caractérisés dans le chapitre 3 consacré au génotypage.

La **figure 2.7** reprend les différentes étapes de la mise en évidence du polymorphisme microsatellite au sein de notre banque d'unigènes.

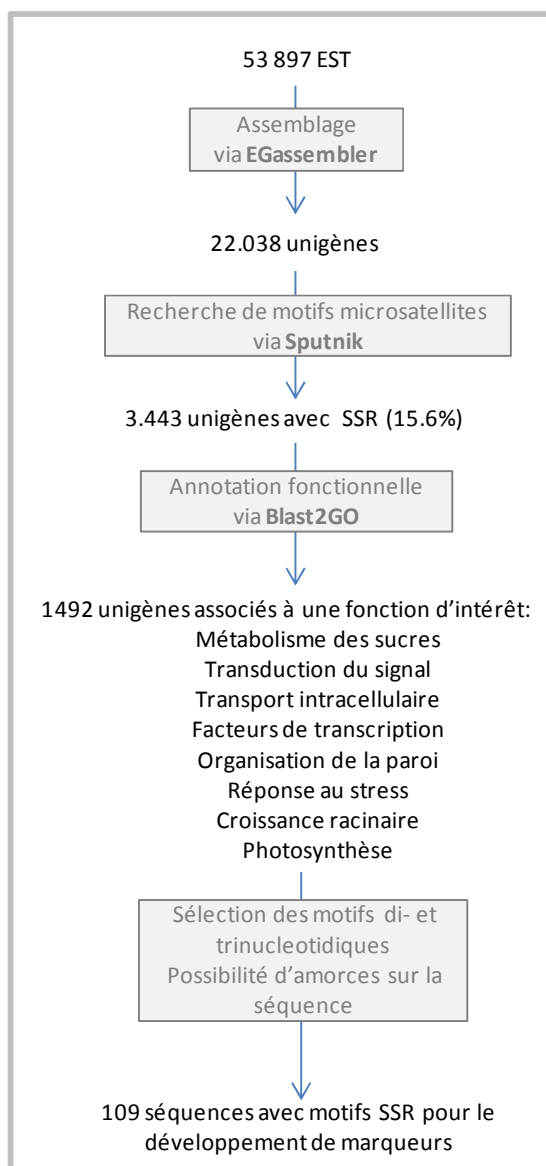


Figure 2.7: Schéma récapitulatif des différentes étapes de mise en évidence du polymorphisme microsatellite au sein de séquences EST

II.III. MISE EN ÉVIDENCE DE SNP

II.III.1. Introduction

Le polymorphisme SNP est le type de polymorphisme le plus fréquemment rencontré au sein d'un génome. Cette grande variabilité due aux mutations ponctuelles est une source importante de marqueurs moléculaires. Les banques EST permettent de mettre en évidence ce polymorphisme à l'aide d'outils bioinformatiques.

Dans cette section, les SNP présents au sein des EST de chicorée de la banque Phytomol ont été identifiés et caractérisés à l'aide d'un pipeline bioinformatique développé par le Compositea Genome Project. Contrairement aux motifs microsatellites, ce n'est qu'en comparant plusieurs séquences similaires entre elles qu'il est possible de détecter le polymorphisme SNP. Il était donc nécessaire, dans un premier temps, de regrouper les séquences identiques au sein d'un contig. Ensuite, les séquences d'un même contig ont été alignées afin de voir les différences au niveau nucléotidique. Enfin, ces différences ont été répertoriées et les éventuelles erreurs de séquençage ont été supprimées.

Les SNP ainsi mis en évidence ont été utilisés pour le développement de marqueurs moléculaires et pour le génotypage (chapitre 3).

II.III.2. Matériel et méthodes

II.III.2.1. Banque de données

La recherche du polymorphisme SNP a été réalisée à partir de la banque EST Phytomol soit 12.226 EST. Les séquences ont été rassemblées dans un fichier en format fasta.

II.III.2.2. Analyses bioinformatiques

La recherche du polymorphisme SNP a été effectuée à partir du protocole Python SDP (Python SNP Discovery Pipeline) (Kozik 2007) disponible sur http://cgpdb.ucdavis.edu/SNP_Discovery_CDS/. Les scripts de chaque programme sont disponibles sur le site.

- Le script Cap3 regroupe dans un contig toutes les séquences ayant au moins un chevauchement de 80 nucléotides avec un minimum de 90% d'identité. Au final, on obtient :
 - Un fichier « .contigs » avec les séquences de chaque contig
 - Un fichier « .singlets » regroupant les EST présentes une seule fois
- Le script ContigExtractor réalise l'alignement des séquences d'un contig.
- Le script MMFinder extrait les SNP et Indels présents au sein de tous les alignements. Deux fichiers sont créés :
 - Un fichier « .MM » regroupant tous les SNP et Indels sans aucune sélection
 - Un fichier « .MMgood » regroupant les SNP et Indels présents au minimum deux fois dans l'alignement
- Le script Contig Viewer permet de visualiser chaque alignement ainsi que la position des SNP et Indels présents au sein du contig.

Chaque contig possédant des SNP-Indels répertoriés dans le fichier « MMgood » a été analysé manuellement. Seuls les SNP répondant aux trois critères suivants ont été retenus :

- Présents plus de deux fois dans l'alignement
- Positionnés à plus de 100 pb des extrémités de la séquence consensus
- Situés en dehors des zones présentant un grand nombre d'erreurs de séquençage

II.III.3. Résultats et discussion

II.III.3.1. Choix de la banque EST

Nous avons cherché à mettre en évidence le polymorphisme SNP présent au sein de la chicorée industrielle, *C. intybus* var. *sativum*. La variété (*sylvestre*, *sativum*, *foliusum* ou *radicchio*) des individus utilisés pour la création des banques EST D et L n'est pas précisée. Afin de sélectionner du polymorphisme intravariétal et non intervariétal, les analyses ont donc été réalisées uniquement sur la banque EST Phytomol. En effet, le polymorphisme entre variétés différentes peut ne pas être présent au sein même d'une variété. Vu que nos

recherches étaient focalisées sur la variété *sativum*, c'est au sein de la banque Phytomol que nous avons recherché des SNP. Pour la recherche de motifs SSR, nous n'avons pas fait la distinction entre les banques EST car nous avons mis en évidence la présence de motifs SSR et non la présence de polymorphisme SSR.

II.III.3.2. Formation et caractérisation des contigs

Afin de regrouper les séquences similaires au sein d'un contig pour réaliser un alignement, deux scripts provenant du pipeline bioinformatique du Composita Genome Project ont été utilisés. L'utilisation des scripts Cap3 et Contig Extractor a permis de générer un jeu de 1.546 contigs. Au total, 43% des EST se retrouvaient donc en plusieurs exemplaires dans la base de données initiale. Ce qui signifie que la plupart des EST étaient en copie unique (53%). Cela peut s'expliquer par le choix restrictif des critères de création de contigs, c'est-à-dire un chevauchement minimum de 80 nt et surtout une identité de 90%.

La taille moyenne était de 619 pb pour les séquences EST brutes et de 781 pb pour les séquences consensus obtenues lors de l'alignement des contigs. Cela représente une augmentation de plus de 25% par rapport aux séquences initiales. Le fait d'obtenir des séquences plus grandes est un avantage, les zones propices à la conception de sondes nécessaires pour le génotypage étant plus nombreuses. C'est à partir de l'alignement des séquences contenues dans ces contigs que le polymorphisme SNP a été mis en évidence.

II.III.3.3. Analyse des SNP

- Stratégie

Afin de mettre en évidence les SNP/Indels présents dans les contigs, le script MM Finder a été utilisé. Trois sélections consécutives ont été réalisées avec des exigences de plus en plus grandes. La première, la sélection MM, reprenait tous les SNP et Indels présents au sein des contigs. Les probabilités de mettre en évidence des erreurs de séquençage et non du polymorphisme réel étaient donc élevées. La sélection MMgood réalisée automatiquement par le programme ne mettait en évidence que les SNP/Indels présents deux fois au sein des contigs. Cette deuxième analyse a permis de diminuer les chances de

sélectionner des erreurs de séquençage. La troisième sélection a été réalisée manuellement. Chaque contig a été visualisé et analysé afin d'estimer si les SNP présents pouvaient être dus à des erreurs de séquençage. Seuls les SNP présents plus de deux fois dans l'alignement, à plus de 100 pb des extrémités de la séquence, et situés en dehors des zones présentant de nombreuses erreurs de séquençage ont été retenus. Au final, les SNP sélectionnés avaient donc de grandes chances d'être de réelles mutations et non des artefacts.

Les séquences paralogues, de par leur grande similarité, pouvaient être regroupées dans un même contig. Cela générerait deux groupes de séquences se distinguant par un grand nombre de SNP. Ce type de profil a donc également été éliminé afin d'éviter le développement de marqueurs amplifiant des familles multigéniques.

- Analyse des résultats

Seuls les SNP présents dans minimum trois séquences au sein d'un contig ont été retenus lors de l'analyse manuelle. Cela signifie qu'il fallait au moins quatre séquences au sein d'un contig pour pouvoir y sélectionner un SNP. Le nombre moyen de séquences par contig était de 3,46. La **figure 2.8** représente la répartition du nombre de séquences par contig.

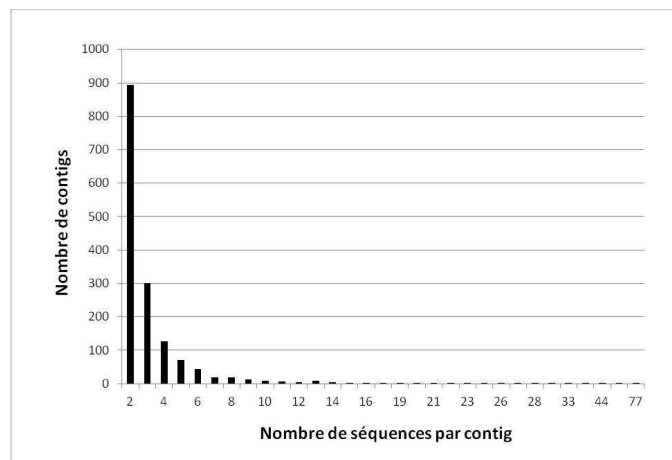


Figure 2.8 : Répartition du nombre de séquences par contig formé à partir de la banque EST Phytomol à l'aide du script Cap3. Les séquences au sein d'un même contig possèdent un chevauchement d'au moins 80 nucléotides avec au minimum 90% d'identité

La majorité des contigs (77%) possédait moins de quatre séquences. Ceux-là ne pouvaient donc pas, suivant nos critères de sélection de SNP, être utilisés. Il nous restait donc 353 contigs potentiellement porteurs de polymorphisme à analyser.

Au final, 294 SNP au sein de 64 contigs ont été retenus. En moyenne, nous avons donc 4,6 SNP par contig, ce qui représente un SNP toutes les 198 paires de bases, autrement dit, 0,5 SNP/100 pb. Cette fréquence laisse donc supposer que des SNP étaient potentiellement présents dans toutes les séquences EST (taille moyenne : 618 pb). Cette fréquence d'occurrence fait donc des SNP des marqueurs de choix pour la cartographie de séquences. Des résultats comparables (entre 0,017 et 0,585 SNP/100 pb) ont été obtenus lors de l'analyse intraspécifique de séquences EST issues de différentes espèces du genre *Solanum* (Jimenez-Gomez and Maloof 2009).

Lors de notre analyse, nous avons tenté d'identifier les régions dans lesquelles se trouvaient les SNP retenus. Les séquences EST de la banque Phytomol ont été séquencées à partir de l'extrémité correspondant au polyA de l'ARN_m (Dauchot et al. 2009). Chaque EST est donc composée de la région 3'-UTR et d'une partie de la région codante du gène. La recherche d'ORF au sein de ces séquences a permis de délimiter la région codante. Au sein de cette région, nous avons caractérisé le type de substitution (synonyme, non-synonyme, non-sens, faux-sens).

La localisation des SNP dans la banque EST Phytomol (**figure 2.9**) a montré que 85% des substitutions dans les régions codantes étaient des mutations synonymes et 15% étaient des mutations non-synonymes. Les mutations synonymes sont plus fréquentes dans le codant que les mutations non-synonymes puisqu'elles ne modifient pas la séquence en acides aminés de la protéine et donc subissent moins de pression de sélection. Notre analyse a également montré qu'il y avait moins de substitutions dans la région 3'UTR que de mutations synonymes dans le codant. Cela peut s'expliquer par le fait que les régions UTR jouent un rôle important dans la régulation de la traduction vu la présence de signaux au sein de ces séquences. Ces mutations semblent avoir un impact plus important sur l'organisme que les mutations synonymes. La pression de sélection est donc vraisemblablement plus forte au sein de ces régions régulatrices.

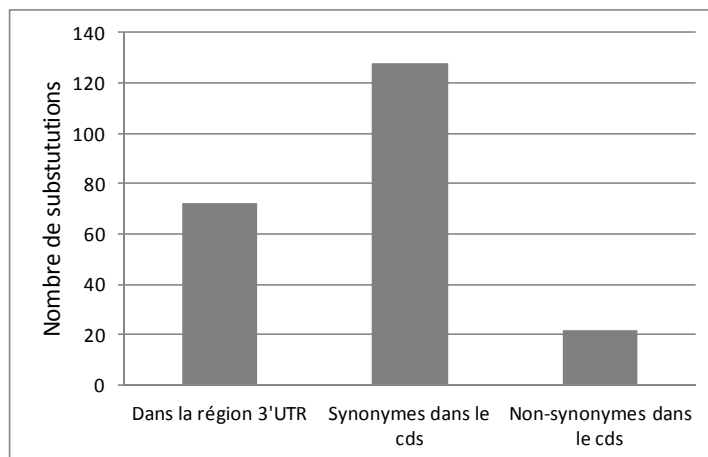


Figure 2.9 : Nombre de substitutions observées dans les régions 3'UTR et dans les régions codantes au sein des séquences EST de la banque Phytomol

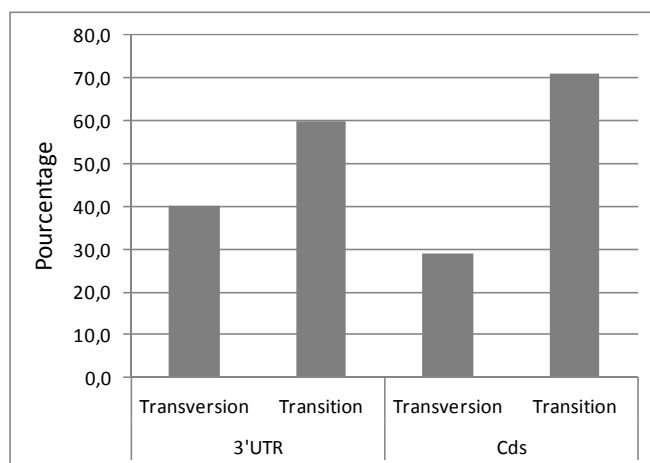


Figure 2.10 : Type de substitution observé dans la région 3'UTR et codante au sein des EST de la banque Phytomol

L'analyse de la nature des mutations dans la région 3'UTR et dans la région codante (**figure 2.10**) a montré que les transitions étaient majoritaires par rapport aux transversions. Cela peut s'expliquer par deux facteurs. Tout d'abord, il semblerait que les mécanismes responsables des transitions soient plus fréquents que les mécanismes responsables des transversions. De plus, les transitions engendrent moins de modifications au niveau des acides aminés que les transversions, ce qui implique une pression de sélection moins

importante sur ce type de mutation (Gibson and Muse 2004). Dans notre cas, les valeurs étaient légèrement différentes entre la région 3'UTR et la région codante où l'on a trouvé respectivement, 60 et 70% de transitions. L'hypothèse selon laquelle les transitions sont plus nombreuses car elles engendrent moins de modifications dans la séquence protéique n'entre pas en ligne de compte pour la région 3'UTR. Cela pourrait donc expliquer cette différence moins marquée entre les deux types de substitutions.

Afin d'évaluer l'impact des mutations non-synonymes dans le codant, nous avons utilisé une table qui évalue la distance physicochimique entre deux acides aminés (Grantham 1974). Cette table se base sur les propriétés telles que la polarité, le volume de la molécule ainsi que la composition chimique. Cette distance permet de quantifier le changement engendré par une mutation au sein de la région codante. La valeur moyenne est de 100. Les substitutions ont été classées avec d'un côté celles possédant une distance physico-chimique inférieure à 100 et de l'autre coté, celles possédant une valeur supérieure à 100 (**figure 2.11**).

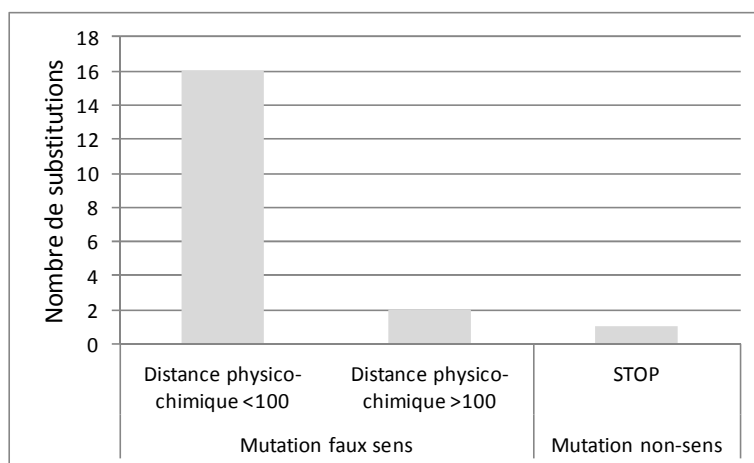


Figure 2.11 : Nombre de substitutions non-sens et faux-sens observées lors de l'analyse des EST de la banque Phytomol. Les mutations faux-sens ont été classées suivant la distance physicochimique (Grantham, 1974) des deux acides aminés (en-dessous ou au-dessus de 100)

88% des mutations faux-sens impliquaient des acides aminés dont la distance physicochimique était inférieure à 100. Cela signifie que ces substitutions engendrent une différence physicochimique au sein de

la protéine plus faible que la moyenne des différences pouvant être induites par une mutation. Cela peut facilement s'expliquer par le fait que des changements mineurs dans la nature de l'acide aminé ont moins d'impact sur la structure de la protéine et donc sur sa fonction. Ce type de mutation subit donc moins de pression de sélection par rapport à une mutation qui implique de grands changements physicochimiques au sein de la protéine.

Les mutations non-sens ont un impact majeur sur la protéine puisqu'elles engendrent une protéine tronquée. Dans notre analyse, seuls 5% des mutations non-synonymes étaient de cette nature.

II.IV. CONCLUSION SUR L'ANALYSE DES BANQUES EST

Les banques EST sont une source inestimable d'informations génétiques pour une espèce non-modèle. Depuis quelques années, des banques EST de chicorée sont apparues : au total, plus de 50.000 EST peuvent être utilisées pour la recherche de polymorphisme à l'aide d'outils bioinformatiques. Le polymorphisme mis en évidence étant associé à une séquence transcrite et donc à une fonction, les marqueurs moléculaires ainsi développés peuvent être directement ou indirectement liés à un phénotype d'intérêt.

Lors de notre analyse, les EST ont été utilisées pour mettre en évidence du polymorphisme microsatellite et SNP : 15% des séquences EST possédaient des motifs SSR et on a observé une fréquence de 1 SNP tous les 194 pb dans les contigs retenus.

Comme attendu, la majorité (59%) des SSR mis en évidence était de type trinuécléotidique dont 50% étaient composés des motifs ATC et AAG. Les SSR dinuécléotidiques sont arrivés en deuxième position. Le motif AG représentait 58% de ce type de microsatellite.

Une sélection basée sur la fonction associée à ces séquences a été réalisée afin d'obtenir des marqueurs liés à une fonction d'intérêt pour la sélection de la chicorée. Ensuite, seuls les SSR de type tri- et dinuécléotidique de taille supérieure à 12 pb ont été retenus. Puisque le taux de mutation est lié au type de microsatellite ainsi qu'à sa taille, ces critères permettaient d'augmenter les chances d'obtenir des marqueurs SSR polymorphes. Au final, 109 EST ont ainsi été retenues pour le développement de marqueurs moléculaires.

Concernant la mise en évidence des SNP, l'utilisation et la maîtrise d'un pipeline bioinformatique ainsi que l'analyse manuelle systématique de chaque contig ont permis de mettre en évidence 294 sites polymorphes contenus dans 64 contigs. La distribution des SNP au sein des régions transcrites (régions 3'UTR et codantes) a montré une variation qui est fonction de la pression de sélection causée par ces mutations. 58% des substitutions étaient des mutations synonymes, 32% étaient des mutations dans la région 3'UTR, 8,5% étaient des mutations faux-sens et 0,5% était des mutations non-sens. La nature et la position des mutations permettent de mieux comprendre l'impact de ces mutations sur le phénotype.

Ces sites polymorphes ont été utilisés pour développer des marqueurs moléculaires. Les méthodes de génotypage mises au point sont décrites dans le chapitre 3. Ces marqueurs ont ensuite été utilisés pour construire la carte génétique de la chicorée (chapitre 4).

III. POLYMORPHISME AU SEIN DE SÉQUENCES ORTHOLOGUES À LA LAITUE

III.I. INTRODUCTION

Cette partie du chapitre est consacrée à la recherche de polymorphisme au sein de séquences de chicorée potentiellement orthologues avec des séquences positionnées sur les cartes génétiques de la laitue.

Le but de cette démarche est d'utiliser le polymorphisme SNP mis en évidence afin de positionner ces séquences de chicorée sur la carte génétique et au final, de pouvoir comparer leurs positions par rapport à leurs orthologues potentielle sur les carte de la laitue.

La cartographie comparée permet d'analyser la synténie et de transférer éventuellement des données génétiques d'une espèce à l'autre. Dans le cas de la chicorée, ce type d'analyse permet d'ouvrir de nouvelles perspectives puisque les ressources génétiques sont relativement peu nombreuses par rapport à d'autres espèces proches comme la laitue. En effet, dans le cadre du Compositea Genome Project, la laitue est une espèce dont les ressources génétiques sont en continuelle croissance. Plusieurs cartes génétiques sont déjà

disponibles (Truco et al. 2007) et la carte physique est en cours de développement (Koozik et al. 2011).

Cette démarche a nécessité dans un premier temps de rechercher des orthologues entre les EST de la laitue déjà cartographiées et les EST de chicorée. Ensuite, les EST de chicorée ainsi retenues ont été analysées au sein de la population de cartographie SAMCHIC afin de mettre en évidence du polymorphisme SNP pouvant être utilisé pour la cartographie. Ces deux étapes sont décrites dans cette section. La cartographie et la cartographie comparée sont décrites dans les chapitres 4 et 5.

III.II. MATÉRIEL ET MÉTHODES

III.II.1. Recherche de séquences orthologues entre la laitue et la chicorée

III.II.1.1. Banques de données

- **Banque *Lactuca* JMR** : 435 EST cartographiées sur les cartes JMR 1, 2 et 3 (CGPDB).
- **Banque *Lactuca sativa-seriola*** : 27.736 EST provenant de *L. lactuca* et *L. sativa* (CGPDB)
- **Banque *Cichorium* NCBI** : 34.517 EST disponibles sur NCBI

III.II.1.2. Reciprocal best BLAST Hit (RBH)

La banque *Lactuca* JMR a été blastée (tBlastx) contre la banque *Lactuca sativa-seriola*. Le RBH a ensuite été effectué grâce à un tBlastx réciproque entre les résultats obtenus dans le premier BLAST et la banque *Cichorium* NCBI. Cette étape a été réalisée à l'aide du programme Seqtools (<http://www.seqtools.dk>) (Rasmussen 2002-2009).

III.II.2. Mise en évidence de SNP dans les séquences orthologues au sein de la population SAMCHIC

III.II.2.1. Population de cartographie SAMCHIC

La population de cartographie F2 est composée de 247 individus. Cette population a été obtenue grâce au croisement de deux parents

de chicorée industrielle ayant subi deux autofécondations et donc partiellement homozygotes (lignée 97-31-02 et lignée 97-25-04) de morphologies racinaires différentes. Un individu F1 a été autofécondé, générant ainsi la population F2. Ce croisement ainsi que les extractions d'ADN ont été réalisés par la société Cosucra Belgium. Pour cette thèse, nous disposons de l'ADN de 200 individus de la F2 sur les 247 initialement disponibles.

III.II.2.2. Conception d'amorces et amplification

Des amorces ont été définies sur les séquences sélectionnées à l'aide du programme primer3 (Rozen and Skaletsky 2000). Le T_m se situe aux alentours de 60°C et le pourcentage de GC est d'environ 50%.

En absence d'ADN de l'individu F1 en quantité suffisante, l'amplification PCR a été réalisée sur un mélange d'ADN provenant de huit individus de la population F2 à une température d'hybridation de 55°C, 90 secondes d'élongation et durant 40 cycles. Afin de marquer à l'aide d'un chromophore les produits d'amplification, des dUTP marqués (Dyomics #648-34 et #618-34) ont été incorporés au mélange PCR.

III.II.2.3. Evaluation du polymorphisme par Ecotilling

Etape 1 : Dénaturation-hybridation : La totalité du produit PCR a été chauffée à 95°C puis refroidie graduellement jusqu'à 4°C.

Etape 2 : Digestion : 100 ng (dye noir), 75 ng (dye vert) ou 50 ng (dye bleu) du produit d'hybridation (maximum 7 µl) ont été portés à un volume de 10 µl à l'aide d'eau milliQ. Cinq µl du mélange d'enzymes suivant ont été ajoutés : 1,5 µl d'enzyme endo1 (Serial Genetics) diluée 50 fois (à l'aide du tampon de dilution), 1,5 µl de tampon de réaction 10 fois, 2 µl d'eau milliQ. Le tout a été incubé pendant deux heures à 37°C.

Etape 3 : Précipitation de l'ADN : La réaction a été arrêtée en ajoutant 5 µl de Stop solution + 60 µl d'éthanol absolu stocké à -20°C (Stop solution : 2 µl d'acétate de sodium (3 M pH 5,2), 1,6 µl d'eau milliQ, 1 µl de glycogène (20 mg/ml), 0,4 µl d'EDTA (0,5M à pH 8)).

Les échantillons ont été centrifugés une heure à 3.500 rpm à 4°C. Le surnageant a été éliminé. 200 µl d'éthanol 70% stocké à -20°C ont été

ajoutés. Le tout a été centrifugé 10 min à 3.500 rpm. Le surnageant a été éliminé. Un second lavage à l'éthanol 70% a été réalisé. Le surnageant a été éliminé. L'éthanol restant a été supprimé en plaçant les échantillons 30 minutes au Speedvac. Les échantillons ont été resuspendus dans 40 µl de formamide désionisée (ABI, #4311320) contenant 0,2 µl de standard de taille (BioVentures, #1-800-235-8938). Les échantillons ont été analysés au séquenceur capillaire Beckman Coulter (Ceq 8000 Genetic Analysis System).

III.II.2.4. Clonage des produits PCR

- Préparation de cellules thermocompétentes

Le tampon CCMB80 a été réalisé : 10 mM KOAc pH 7,0, 80 mM CaCl₂.2H₂O, 20 mM MnCl₂.2H₂O, 10 mM MgCl₂.6H₂O, 10% glycérol, pH 6,4. Celui-ci a été stérilisé par filtration 0,22 µm et stocké à 4°C.

Une pré-culture au départ d'une colonie isolée obtenue après avoir strié la souche utilisée sur une boîte de milieu LB a été lancée pendant une nuit à 37°C dans 20 ml de milieu LB avec antibiotiques à 180 rpm. Une culture de jour a été réalisée en transférant 5 ml de la pré-culture dans 200 ml de milieu LB avec antibiotiques. L'évolution de la DO à 590 nm a été suivie pour atteindre environ 0,3.

La suite des manipulations a été faite sur glace. Le milieu de culture a été transféré dans 4 tubes Falcon de 50 ml et centrifugé à 4°C, 4.500 rpm pendant 10 minutes. Le surnageant a été éliminé et les culots ont été resuspendus délicatement dans un total de 80 ml de milieu CCMB80 refroidis à 4°C. Les tubes ont été déposés sur glace pendant 20 minutes et centrifugés une nouvelle fois à 4°C, 4.500 rpm pendant 10 minutes. Le culot a été resuspendu dans 8 ml de milieu CCMB80. La densité de cellules a été testée en mélangeant 200 µl de cellules avec 800 µl de milieu LB. Si nécessaire, les cellules peuvent être diluées avec du milieu CCMB80 pour attendre une DO 590 comprise entre 1,0 et 1,5. Les cellules ont été incubées sur glace pendant 20 minutes puis transférées en aliquotes de 50 µl et stockées à -80°C.

- Préparation des vecteurs TA

Le vecteur pBluescript II ks (Stratagen) a été extrait à l'aide du Quiagen Plasmid Maxi kit (#12162) suivant les instructions du fabricant.

La digestion avec une enzyme de restriction coupant à bout droit a été réalisée à l'aide du mélange décrit dans le **tableau 2.1**.

Tableau 2.1 : Mélange réalisé pour la digestion du plasmide pBluescript II ks avec EcoRI

Dans un volume de 50 µl :	Quantité
Plasmide (1 µg/µl)	10 µl
Tampon EcoR V (NEB) 10x	5 µl
Eco R V 10U/µl (NEB)	5 µl

Le tout a été mélangé au vortex, centrifugé quelques secondes et incubé deux heures à 37°C. Le vecteur a été purifié à l'aide d'une colonne Nucléospin (Macherey Nagel). Les extrémités « T » ont été ajoutées au vecteur grâce à la terminal deoxynucleotidyl-transferase (TdT). Le mélange (**tableau 2.2**) a été incubé à 37°C pendant 1h30 et purifié à l'aide d'une colonne Nucléospin (Macherey Nagel).

Tableau 2.2 : Mélange pour l'ajout des extrémités T sur le vecteur pBluescript II ks digéré

	Quantité
Plasmide digéré et purifié	46 µl (+/- 10 µg)
TdT buffer 5X	15 µl
CoCl ₂ 25 mM	7.5 µl
ddITP 1 mM (USB)	1.5 µl
Terminal Transferase (400U/µl) (Roche #03 333 566 001)	1 µl
Eau	4 µl

Afin de sélectionner les vecteur TA, une ligation a été réalisée pour ne garder que les vecteurs non ligués : les vecteurs TA corrects sont incapables de s'autoliguer car le T terminal est un didéoxyribose. Le mélange décrit dans le **tableau 2.3** a été réalisé et incubé à 16°C pendant la nuit.

Tableau 2.3 : Mélange réalisé pour la ligation des vecteurs pBluescript II ks dans le but de sélectionner les vecteur TA

Plasmides purifiés	44 µl
TP ligase 10X	5 µl
T4 DNA ligase (3U/µl) (Promega)	1 µl

Le produit de ligation a été passé sur gel d'agarose et ensuite purifié sur Nucléospin. La concentration du vecteur a été estimée par migration sur gel d'agarose. Le vecteur est aliquoté et stocké à -20°C.

▪ Ligation du vecteur TA et du produit PCR

Le mélange détaillé dans le **tableau 2.4** a été réalisé et incubé à 16°C pendant une nuit :

Tableau 2.4 : Mélange réalisé pour la ligation des produits PCR au sein des vecteur TA

Ligase (400 000 U/ml, NEB)	1 µl
Tampon ligase 10 x	1 µl
Vecteur pBluescript TA (Stratagen) (30 ng/µl)	2 µl
Produit PCR purifié	2 µl
Eau	4 µl

▪ Procédure de transformation

Six µl de produit de ligation ont été ajoutés à 50 µl de cellules compétentes (TOP10F') ; celles-ci ont ensuite été incubées sur glace pendant 30 minutes. Les cellules ont ensuite été soumises à un choc thermique de 30 secondes à 42°C et placées immédiatement sur glace pendant deux minutes. 250 µl de milieu SOC ont été ajoutés et les cellules ont été incubées sous agitation à 37°C pendant 45 à 60 minutes à 225 rpm. 300 µl de bactéries ont été étalés sur une boîte de milieu LB avec sélection et incubés une nuit à 37°C.

III.II.2.5. PCR sur colonie et séquençage

Une PCR sur colonie a été réalisée à l'aide des amorces du **tableau 2.5**.

Tableau 2.5 : Amorces utilisées pour l'amplification du vecteur pBluescript II ks

Reverse primer (pBluescript II ks +)	5'-GGAAACAGCTATGACCATG-3'
Forward primer (pBluescript II ks +)	5'-CATTTTGCTGCCGGTCA-3'

Les produits PCR avec insert ont ensuite été séquencés via la société Macrogen (<http://dna.macrogen.com/eng/>). L'analyse, l'assemblage

et l'alignement des séquences ont été réalisés à l'aide du programme VectorNTI[®] (Invitrogen).

III.III. RÉSULTATS ET DISCUSSION

III.III.1. Recherche de séquences orthologues entre la laitue et la chicorée

La méthode sélectionnée pour établir un lien d'orthologie entre les EST de laitue et de chicorée est la méthode RBH (Reciprocal best Blast Hit). Lors de la réalisation de tBalstx entre les banques EST des deux espèces, seules les séquences se détectant réciproquement comme meilleurs résultats de blast sont définies comme des orthologues potentiels.

L'approche RBH a été réalisée en plusieurs étapes (**figure 2.12**) :

- Les 435 EST positionnées sur les carte JMR 1, 2 et 3 de la laitue (CGPDB) ont été blastées contre 27.736 EST appartenant à *L. sativa* et *L. serriola*. Cette étape a pour but d'identifier les EST cartographiées dans un jeu de données plus large.
- Les 435 EST de *L. sativa* et *L. serriola* ainsi obtenues ont ensuite été blastées contre les 34.517 EST de chicorée disponibles sur NCBI.
- Les 435 séquences de chicorée retenues ont à leur tour été blastées contre la banque *L. sativa-serriola* afin de sélectionner les paires de séquences présentant un RBH.

Au final 230 séquences sur les 435 cartographiées (53 %) présentaient un RBH avec une séquence de chicorée. Cette méthode se base théoriquement sur le fait que l'on dispose des génomes complets des deux espèces. Cependant, les banques EST que nous avons utilisées ne reprennent pas la totalité des séquences codantes. En effet, dans la banque « *Cichorium* NCBI », les unigènes représentent environ 19 Mb, soit 1,3 % du génome total (1.400 Mb, (Koopman 2002)). Si l'on estime que, comme chez *A. thaliana*, la partie codante représente 4 % du génome total (AGI 2000), nos données recouvriraient 32 % du génome codant de la chicorée. Il est donc très important de garder en mémoire que les paires de séquences obtenues lors de l'analyse RBH sont des orthologues potentiels.

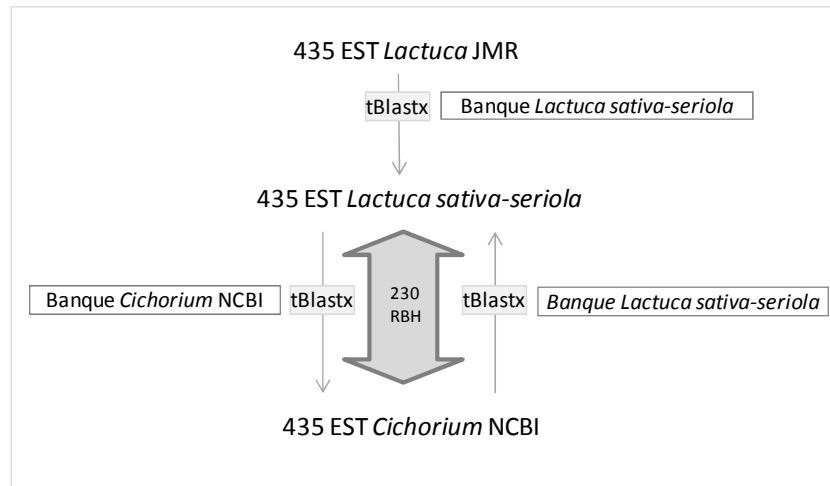


Figure 2.12 : Schéma récapitulatif de l'analyse RBH (Reciprocal Best Blast Hit). 435 EST positionnées sur les cartes génétiques de la laitue (banque EST *Lactuca JMR*) ont été utilisées pour réaliser un tBlastx contre la banque *Lactuca sativa-seriola*. Les 435 best BLAST Hit de la banque *Lactuca sativa-seriola* ont été blastés (tBlastx) contre la banque *Cichorium NCBI*. Les 435 best BLAST Hit obtenues ont été blastées (tBlastx) contre la banque *Lactuca sativa-seriola* générant ainsi 435 best BLAST Hit. Au final, 230 séquences de laitue possèdent un RBH avec une séquence de chicorée provenant de la banque *Cichorium NCBI*

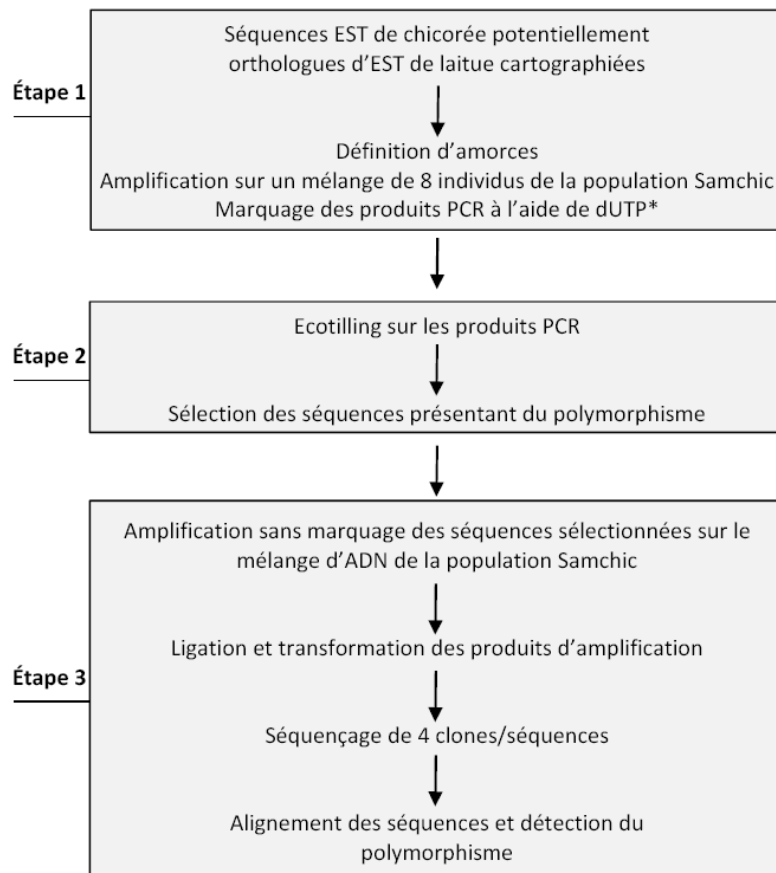


Figure 2.13: Récapitulatif des différentes étapes pour mettre du polymorphisme en évidence au sein de séquences d'intérêts dans la population de ségrégation SAMCHIC

III.III.2. Mise en évidence du polymorphisme SNP contenu dans les séquences orthologues au sein du croisement SAMCHIC

Afin de positionner ces orthologues potentiels sur la carte génétique de la chicorée dans le but de réaliser la cartographie comparée avec la laitue, ces séquences doivent présenter du polymorphisme au sein de la population de ségrégation SAMCHIC développée par la société Cosucra (voir matériel et méthodes point III.II.2.1). Ce paragraphe décrit la stratégie développée pour réaliser cette étape (**figure 2.13**) ainsi que les résultats obtenus.

▪ Etape 1

Des amorces ont été définies pour les 230 EST de chicorée retenues lors de l'analyse RBH. Une amplification a été réalisée sur un mélange d'ADN génomique (ADNg) de huit individus appartenant à la population F2 utilisée pour la cartographie. 92% des amorces testées ont donné une amplification.

Les amorces ont été définies à partir des séquences EST. On pourrait donc s'attendre à ce qu'une plus grande partie des amorces chevauche des régions intron-exon dans l'ADN génomique ou que la présence d'introns de grande taille empêche l'amplification. Il semblerait que dans notre cas, la grande majorité des amorces n'ait pas rencontré ces problèmes.

Les produits PCR ont été marqués via l'incorporation de dUTP couplés à un chromophore. Lors de la PCR, la taq polymérase pouvait incorporer en face d'une adénine soit un dTTP soit un dUTP marqué. Cela a permis d'obtenir des fragments couplés à des chromophores sans devoir utiliser d'amorces marquées pour chaque séquence amplifiée.

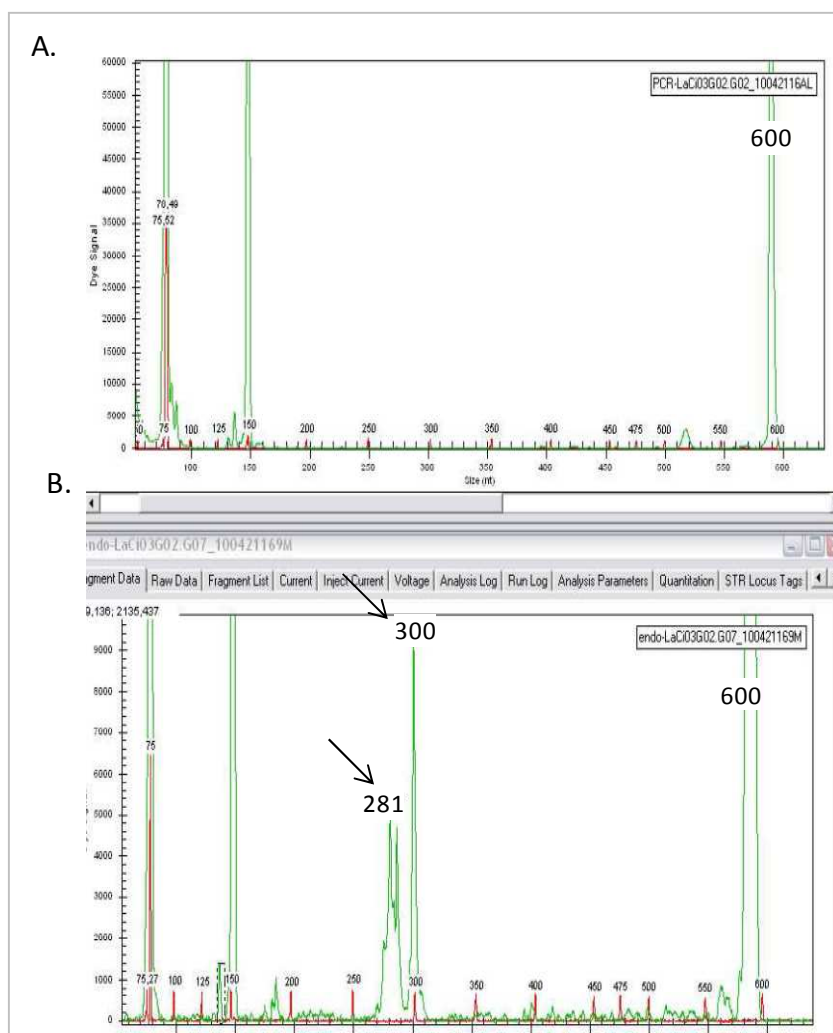


Figure 2.14 : Exemple d'analyse de la taille des fragments obtenus par Ecotilling. Ordonnée : taille des fragments en pb. Abscisse : intensité du signal fluorescent. Rouge : Standard de taille. Vert : fragments analysés. (A) Fragments obtenus après dénaturation et réhybridation. Le fragment de 600 pb représente la séquence totale non digérée. (B) Fragments obtenus après digestion des produits de réhybridation par une endonucléase. Les fragments de 281 pb et de 300 pb représentent la position des sites digérés par l'enzyme et donc présentant du polymorphisme pour les individus analysés

- Etape 2

L'Ecotilling a été réalisé sur ces produits d'amplification. Cette méthode permet de digérer les mésappariements dans l'ADN double brin via une endonucléase. Après une étape de dénaturation, une réhybridation des deux brins d'ADN se fait de manière très lente, ce qui permet à deux brins initialement non complémentaires de s'associer. Si ceux-ci ne sont pas parfaitement identiques, des mésappariements se formeront et seront digérés. La taille des fragments obtenus permet de localiser le polymorphisme présent entre les différents allèles.

L'analyse des 230 séquences en Ecotilling (**figure 2.14**) a montré du polymorphisme pour 65 séquences, soit 28 %. Les parents utilisés pour le croisement SAMCHIC possédaient des caractéristiques morphologiques distinctes au niveau de la racine, mais la distance génétique entre eux n'a pas été calculée. Suivant les observations faites ici, nous pouvons émettre l'hypothèse que les deux parents étaient génétiquement proches, vu le peu de polymorphisme observé au sein de leur descendance. De plus, les séquences que nous avons analysées font partie des régions transcrites du génome. Le polymorphisme au sein de ces régions est, de manière générale, plus faible que dans les régions non transcrites. Ce facteur influence donc également le taux de polymorphisme observé au sein de nos séquences.

- Etape 3

Afin de vérifier ce polymorphisme, les amplicons obtenus via un mélange de huit ADN ont été séquencés. Afin d'analyser un allèle à la fois, chaque produit PCR a été cloné et ce sont les produits de la PCR sur colonie qui ont été séquencés. 65 clonages ont donc été réalisés et en moyenne quatre clones par produit PCR ont été séquencés. L'alignement des différents allèles a permis d'identifier du polymorphisme SNP dans 56 séquences. Certaines séquences présentant beaucoup de polymorphisme au sein d'un même alignement ont été définies comme des paralogues potentiels et ont été éliminées.

Tableau 2.6 : Nom, fonction et provenance de 35 séquences dont le polymorphisme a été analysé au sein de la population SAMCHIC dans le but d'être cartographiées

Séquence cible	Fonction	Provenance de la séquence	Polymorphisme dans le croisement SAMCHIC
PIE	Floraison	Dognaux 2009	Non
Floral organ identity	Floraison	Dognaux 2009	Non
Constants- like (COL)	Floraison	Dognaux 2009	Non
Early-flowering 6 (ELF6)	Floraison	Dognaux 2009	Oui
Vin3-Like 1 (VIL 1)	Floraison	Dognaux 2009	Oui
Early-flowering 4 (ELF 4)	Floraison	Dognaux 2009	Non
Early-flowering 7 (ELF7)	Floraison	Dognaux 2009	Non
Gigantea (GI)	Floraison	Dognaux 2009	Oui
Constants (CO)	Floraison	Dognaux 2009	Oui
Flowering time control protein (FY)	Floraison	Dognaux 2009	Non
Vernalization insensitive 3 (VIN3)	Floraison	Dognaux 2009	Oui
Vernalization independence 4 (VIP4)	Floraison	Dognaux 2009	Non
Supressor of Frigida 4 (SUF-4)	Floraison	Dognaux 2009	Non
Flowering Locus C (FLC)	Floraison	Dognaux 2009	Non
AY153822.1, resistance protein candidate gene	Pathogenèse	Hespeels 2009	Oui
AY153825.1, resistance protein candidate gene	Pathogenèse	Hespeels 2009	Non
Chryptochrome 2 (CRY2)	Floraison	Dognaux 2009	Oui
Pseudo-Reponse Regulator 5 (APRR5)	Floraison	Dognaux 2009	Non
Early-flowering 8 (ELF8)	Floraison	Dognaux 2009	Non
Séquence avec domaine TIR-NBS-LRR	Pathogenèse	Hespeels 2009	Non
CHICORY WAK LIKE 1	Pathogenèse	Hespeels 2009	Oui
CHICO WAK LIKE 2	Pathogenèse	Hespeels 2009	Oui
CHICO WAK LIKE 3	Pathogenèse	Hespeels 2009	Oui
contig1531	fructose-bisphosphate aldolase	Banque Phytomol	Oui
contig605	transaldolase	Banque Phytomol	Non
contig1	poly-binding protein	Banque Phytomol	Oui
contig578	-	Banque Phytomol	Non
contig112	ubiquitin conjugating enzyme	Banque Phytomol	Oui
contig1130	-	Banque Phytomol	Non
contig1156	udp-glucose pyrophosphorylase	Banque Phytomol	Non
contig1248	polypeptide of oxygen-evolving complex	Banque Phytomol	Non
contig1273	40s ribosomal protein	Banque Phytomol	Non
contig1359	ribulose bisphosphate carboxylase small chain 2b precursor	Banque Phytomol	Oui
contig1475	ribulose-bisphosphate carboxylase oxygenase small subunit protein	Banque Phytomol	Non
contig1501	protein kinase	Banque Phytomol	Oui

Au final, 183 SNP ont été retenus dans 56 séquences, ce qui représente une moyenne de 3,2 SNP par séquence, soit une fréquence de un SNP tous les 224 pb. Cette valeur est du même ordre de grandeur que celle obtenue lors de l'analyse de la banque Phytomol (un SNP /198 pb) (chapitre 2 point I.IV.3.3). Lors de l'analyse Phytomol, le polymorphisme SNP a été identifié entre 10 individus provenant du même cultivar. Concernant les SNP analysés ici, le polymorphisme mis en évidence concerne les variations présentes entre les deux parents du croisement de la F2. Il semblerait que la fréquence des SNP soit similaire entre les deux collections.

IV. POLYMORPHISME AU SEIN DE SÉQUENCES SUPPLÉMENTAIRES

Afin d'augmenter le nombre de marqueurs sur la carte génétique, le polymorphisme de séquences supplémentaires a été analysé au sein de la population SAMCHIC.

IV.I. CHOIX DES SÉQUENCES

- Séquences provenant d'une carte génétique de chicorée

Une carte consensus de chicorée a été publiée en 2010 (Cadalen et al. 2010). Cette carte résulte de l'intégration de trois cartes génétiques de chicorée industrielle et de chicoré witloof. Parmi les marqueurs cartographiés, 46 STS (Sequence Tagged Sites) provenant d'EST, d'ADNc, de gènes d'intérêt et de séquences non codantes ont été analysés par SSCP (Single Strand Conformation Polymorphism) et positionnés sur la carte. Les paires d'amorces utilisées pour amplifier les STS sont disponibles dans l'article. Afin de faire le lien entre notre carte et celle de Cadalen, nous avons tenté de mettre en évidence du polymorphisme dans ces 46 séquences au sein de notre croisement dans le but de les cartographier.

- Séquences associées à la réponse aux pathogènes et à la floraison

Afin d'augmenter le nombre de séquences à cartographier, nous avons également analysé le polymorphisme de séquences d'intérêt déjà analysées au laboratoire (Dognaux 2009; Hespels 2009). 35

Tableau 2.7 : Récapitulatif de la recherche de polymorphismes SSR et SNP au sein des banques EST, des séquences orthologues à la laitue et de séquences d'intérêts chez la chicorée

Type de marqueurs	Données sources	Critères de selection	Nombre de séquences retenues avec un locus SSR ou SNP	Nombre total de sites SSR ou SNP retenus	Fréquence des mutations	
SSR	22.038 unigènes provenant des banques EST Phytomol, L, et D	Caractéristiques du microsatellite, fonction associée à la séquence	109	109	-	-
SNP	Banque de 12.226 EST (Phytomol)	SNP présents 3x dans le contig, qualité de la séquence	64	294	1 SNP/198 pb	4,6 SNP/séquence
SNP	230 EST orthologues aux EST cartographiées chez la laitue	SNP présent dans la population Samchic	56	183	1 SNP/224 pb	3,2 SNP/séquence
SNP	85 séquences d'interets (carte génétique Cadalen, pathogénèses, floraison, ...)	SNP présent dans la population Samchic	38	107	1 SNP/199 pb	2,8 SNP/séquence
Total	-	-	267	693	1 SNP/207 pb	3,5 SNP/séquence

séquences dont la fonction est associée entre autres à la réponse aux pathogènes ainsi qu'à la floraison ont donc été également ajoutées à notre jeu de données (**tableau 2.6**).

IV.II. ANALYSE DU POLYMORPHISME DES SÉQUENCES

Au total, le polymorphisme de 81 séquences a été analysé suivant la démarche expliquée dans la **figure 2.13** décrite dans le point III ci-dessus. 38 séquences soit 47 % présentaient du polymorphisme SNP au sein du croisement. 107 SNP ont été mis en évidence représentant 2,8 SNP par séquence. Si l'on regarde par rapport à la taille totale des séquences analysées, cela nous donne une fréquence de un SNP tous les 199 pb.

V. CONCLUSION DU CHAPITRE

Dans cette étape du travail, nous avons recherché au sein des données disponibles pour la chicorée, du polymorphisme de type microsatellite et SNP afin de développer des marqueurs moléculaires pouvant être utilisés pour la construction d'une carte génétique à l'aide d'une population F2 nommée SAMCHIC. Le **tableau 2.7** résume les différents résultats obtenus.

Dans un premier temps, nous avons mis en évidence du polymorphisme SSR au sein de banques d'unigènes à l'aide d'outils bioinformatiques. 15% de ces séquences contenaient un motif microsatellite de plus de 10 pb. Suite à la sélection des séquences suivant leur fonction ainsi que suivant la nature et la taille des SSR, un jeu de 109 séquences a été retenu afin d'être utilisé pour le développement de marqueurs.

La banque Phytomol a été utilisée afin de générer des alignements permettant de mettre en évidence des mutations ponctuelles. Seuls les SNP présents au moins trois fois dans un contig et localisés dans des régions ayant peu d'erreurs de séquençage ont été retenus. Au total, 294 SNP présents dans 64 contigs ont été sélectionnés.

Les sites SSR et SNP ont été analysés afin de voir la distribution de ces locus suivant leur nature. Concernant les microsatellites, 59 % étaient des trinuécléotides, ce qui est cohérent puisque les mutations au sein de ces motifs ne modifient pas le cadre de lecture dans la séquence.

Chez les SNP, les transitions étaient plus fréquentes que les transversions, que ce soit dans la région 3'-UTR (60-40%) ou dans les régions codantes (70-30%). Ce phénomène est rencontré dans la plupart des génomes et cela peut s'expliquer par l'impact moins important des transitions sur le code génétique et également par le type de mécanisme impliqué dans ces mutations.

Au niveau de la localisation des mutations ponctuelles, on a observé dans la banque Phytomol, une distribution en concordance avec la théorie : les mutations synonymes sont les plus nombreuses, suivies par les mutations dans la région 3'-UTR, puis par les mutations faux-sens et enfin, en dernier lieu, les mutations non-sens.

Au final, suite à l'analyse des banques EST, 173 séquences possédant des motifs SSR et SNP ont été retenues.

Afin de réaliser une cartographie comparée entre la chicorée et la laitue, le polymorphisme au sein de séquences orthologues à la laitue a été recherché chez huit individus appartenant à la F2 du croisement SAMCHIC. Au total, 56 EST de chicorée orthologues d'EST cartographiées chez la laitue ont montré du polymorphisme de type SNP au sein de la population de cartographie. Avec la même démarche, 38 séquences d'intérêt ont montré du polymorphisme au sein de la population F2 et peuvent elles aussi être potentiellement positionnées sur la carte. Au total, 94 séquences présentant du polymorphisme dans la population SAMCHIC ont ainsi été retenues.

Au final, 693 motifs SSR et SNP ont été mis en évidence au sein de 267 séquences. Celles-ci ont été utilisées pour développer des marqueurs moléculaires efficaces et faciles d'utilisation (chapitre 3). Les marqueurs valides ont ensuite servi pour la construction d'une carte génétique (chapitre 4) et pour la cartographie comparée avec la laitue (chapitre 5)

Chapitre 3

Génotypage

Chapitre 3. GÉNOTYPAGE

I. BUT DU CHAPITRE

Dans le chapitre précédent, du polymorphisme a été mis en évidence au sein de séquences transcrites. Ce polymorphisme a été utilisé dans le présent chapitre pour la création de marqueurs géniques, co-dominants et spécifiques de locus afin de les utiliser pour la construction de la carte génétique. Cependant, vu l'origine de ces marqueurs, nous nous attendions à observer moins de polymorphisme que pour des marqueurs non-géniques dans la population de ségrégation. Afin de contrer ce problème, nous nous sommes focalisés dans un premier temps sur l'utilisation de marqueurs AFLP afin de générer rapidement un grand nombre de marqueurs polymorphes pour la cartographie. Dès qu'un nombre suffisant de marqueurs AFLP a été obtenu, des marqueurs géniques ont été ajoutés. Comme attendu, ces marqueurs présentaient peu de polymorphisme dans la population de ségrégation. Afin de développer un maximum de marqueurs, nous avons donc diversifié la nature des sites à analyser. Dans un premier temps les SSR ont été utilisés car ils sont en règle générale très polymorphes. Ensuite, nous nous sommes concentrés sur les marqueurs SNP car ils sont plus fréquents au sein du génome.

Dans ce chapitre, nous parlerons tout d'abord de l'utilisation des marqueurs AFLP, ensuite nous parlerons des résultats obtenus à l'aide des marqueurs SSR ; enfin, le développement de marqueurs SNP sera détaillé.

II. GÉNOTYPAGE À L'AIDE DES MARQUEURS AFLP

II.1. INTRODUCTION

Afin d'obtenir un grand nombre de marqueurs pour le génotypage de la population de ségrégation SAMCHIC, les marqueurs AFLP ont été choisis. Il est en effet possible d'obtenir rapidement un grand nombre de marqueurs AFLP sans connaître au préalable la position ou la nature du polymorphisme. Des banque EST n'ayant pas toujours été disponibles chez la chicorée, cette méthode s'est avérée très intéressante pour saturer la carte génétique et ainsi construire une carte framework sur laquelle des marqueurs co-dominants ont été ajoutés.

Afin de réaliser cette méthode de génotypage (**figure 1.10**), chapitre 1), l'ADN génomique a été clivé par deux enzymes de restriction (EcoRI et MseI). Des adaptateurs de séquences connues ont été ajoutés aux extrémités des fragments de restriction, générant ainsi une matrice pour l'amplification. Une pré-amplification a été réalisée à l'aide d'amorces de séquences complémentaires à la séquence des adaptateurs et des sites de restriction. Ces amorces sont prolongées en 3' par une base arbitraire. La deuxième amplification, dite sélective, a été faite à l'aide d'amorces identiques aux premières mais prolongées à l'extrémité 3' par deux ou trois nucléotides arbitraires. Ces amorces sélectives ont permis de réduire le nombre de fragments amplifiés à une centaine. Ces fragments ont été séparés et détectés par électrophorèse capillaire grâce au marquage fluorescent réalisé lors de l'amplification sélective.

Cette étape de la cartographie a été réalisée de façon conjointe avec l'UCL (Louvain-la-Neuve) par Claire-Noël Thienpont de l'Unité d'Ecophysiologie et d'Amélioration végétale (ECAV) du professeur X. Draye. Les analyses réalisées par l'UCL ont été faites sur les 247 individus de la population SAMCHIC et celles réalisées dans ce travail ont été faites sur 200 individus de la population de ségrégation.

Dans ce chapitre, les résultats obtenus avec plusieurs combinaisons d'amorces pour le génotypage de la population SAMCHIC sont décrits. Ensuite, les problèmes rencontrés lors de l'analyse des données sont détaillés. Enfin, nos résultats sont comparés avec ceux obtenus par

l'ECAV afin d'identifier les conditions optimales pour l'utilisation des marqueurs AFLP.

II.II. MATÉRIEL ET MÉTHODES

II.II.1. Ligation des deux brins des adaptateurs

Deux tubes contenant les mélanges suivants ont été préparés à l'aide des oligos du **tableau 3.1** :

- pour l'adaptateur EcoRI (concentration finale 50 µM) :
 - 100 µl d'EcoRI link1 (100 µM)
 - 100 µl de EcoRI link2 (100 µM)
- pour l'adaptateur MseI (concentration finale 50 µM)
 - 100 µl de MseI link 1 (100 µM)
 - 100 µl de MseI link2 (100 µM)

Tableau 3.1 : Oligos utilisés pour la formation des adaptateurs EcoRI et MseI

EcoRI link1	5'-CTCGTAGACTGCGTACC-3'
EcoRI link 2	5'-AATTGGTACGCAGTCTAC-3'
MseI link 1	5'-GACGATGAGTCCTGAG-3'
MseI link 2	5'-TACTCAGGACTCAT-3'

Les tubes ont été mis pendant cinq minutes à 95°C et ensuite refroidis progressivement jusqu'à atteindre la température ambiante. Les adaptateurs EcoRI ont été dilués 10 x avec du tampon TE (Tris 10 mM, EDTA 0,1 mM, pH 8) afin d'avoir une concentration finale de 5 µM.

II.II.2. Restriction et ligation des adaptateurs

La digestion à l'aide des enzymes de restriction et la ligation des adaptateurs ont été réalisées en même temps à partir du mélange et du programme décrits dans les **tableaux 3.2 et 3.3**.

Tableau 3.2 : Mélange réalisé pour l'étape de la restriction-ligation de l'AFLP

Réactifs	Concentration initiale	Quantité finale	Volume/échantillon (µl)
T4 DNA Ligase Buffer with ATP (New England Biolabs)	10X	1X	0,55
NaCl	0,5 M	0,05 M	0,55
BSA (Sigma)	1 mg/ml	0,275 µg	0,275
Mse I	10 U/µl	0,5 U	0,05
EcoR I	10 U/µl	2,5 U	0,25
Mse I Adaptateur	50 µM	5 µM	0,5
EcoR I Adaptateur	5 µM	0,5 µM	0,5
Ligase T4 DNA (New England Biolabs)	2 000 U/µl	100 U	0,05
Eau milliQ stérile			0,75
ADNg	50 ng/µl	100 ng	2
Volume total :			5,525

Tableau 3.3 : Programme réalisé pour l'étape de la restriction-ligation de l'AFLP

Réaction de Restriction-Ligation		
Etapes	Température (°C)	Durée
Restriction	37	2 heures 30 minutes
Ligation	20	3 heures
Ligation	16	Infini

II.II.3. Amplification pré-sélective

Le produit de la restriction-ligation a été dilué 10 x dans du tampon Tris 10 mM à pH 8,0. L'amplification pré-sélective a ensuite été réalisée à l'aide du mélange et du programme PCR détaillés dans les **tableaux 3.4 et 3.5**.

Tableau 3.4 : Mélange réalisé pour l'amplification pré-sélective de l'AFLP

Réactifs	Volume / échantillon (µl)
AFLP EcoRI et MseI Preselective Amplification	0,5
AFLP Amplification Core Mix (Applied Biosystems)	3,75
Produit de restriction-ligation dilué 10 X	2
Volume total :	6, 25

Tableau 3.5 : Programme PCR pour l'amplification pré-sélective de l'AFLP

Réaction de pré-amplification		
Température (C°)	Durée (minutes)	Répétition
94	02 :00	1 X
94	00 :20	20 X
56	00 :30	
76	02 :00	
60	30 :00	

II.II.4. Amplification sélective

Les 12 combinaisons possibles entre E32, E35, E33 et M47, M56, M57 et M59 ont été testées à l'aide du mélange et du programme PCR décrits dans les **tableaux 3.6** et **3.7**.

Le code des différentes amorces peut être consulté sur <http://wheat.pw.usda.gov/ggpages/keygeneAFLPs.html>.

Tableau 3.6 : Mélange réalisé pour l'amplification sélective de l'AFLP

Réactifs	Concentration initiale	Concentration finale	Volume (µl)
Primer MseI-CNN (Eurogentec)	5 µM	0,25 µM	0,25
Primer D4-EcoRI-ANN (Proligo)	1 µM	0,05 µM	0,25
AFLP Amplification Core Mix			3,75
Produit d'amplification présélective dilué 10x			0,75
Volume total :			5

Tableau 3.7 : Programme PCR pour l'amplification pré-sélective de l'AFLP

Répétition	Température (°C)	Durée (minutes)
1X	94	02 :00
10X	94	00 :20
	66	00 :30
	Diminuer la température après 1 cycle de 1°C par cycle	
	72	02 :00
20X	94	00 :20
	56	00 :30
	72	02 :00
1X	60	30 :00

II.II.5. Migration des fragments AFLP au séquenceur

Les produits d'amplification ont été préparés pour le passage au séquenceur à l'aide du mélange décrit dans le **tableau 3.8**.

Tableau 3.8 : Mélange réalisé pour la migration des fragments AFLP sur séquenceur capillaire Beckman Coulter

Réactifs	Volume (µl)
SLS (Beckman) ou Formamide désionisée (Applied Biosystem)	40
Standard de taille 600 (Beckman #608095)	0,2
Produit d'amplification sélective	2

Les échantillons ont été analysés à l'aide du séquenceur capillaire Beckman Coulter (Ceq 8000 Genetic Analysis System).

II.II.6. Analyse des sorties AFLP

Chaque sortie AFLP a été analysée automatiquement suivant les critères décrits dans le **tableau 3.9**. Chaque sortie a ensuite été analysée manuellement.

Tableau 3.9 : Paramètres d'analyse des sorties AFLP suivant la paire d'amorces utilisée lors de l'amplification sélective pour le programme CEO 8000 Genetic Analysis System (Beckman Coulter)

Combinaison d'amorces	Slope threshold	Relative peak height threshold	Model
E32M56	10	25	Quartic
E32M57	10	25	Quartic
E35M47	10	35	Quartic
E35M56	10	35	Quartic
E35M57	10	35	Quartic
E35M59	10	25	Quartic
Autres	10	25	Quartic

II.III. RÉSULTATS ET DISCUSSION

II.III.1. Choix des amorces pour l'amplification sélective

Afin d'optimiser les résultats AFLP, douze combinaisons d'amorces pour l'amplification sélective ont été testées (**tableau 3.10**). Parmi celles-ci, quatre ont été sélectionnées car elles présentaient des résultats optimaux par rapport au nombre de fragments amplifiés et par rapport à l'intensité du signal chez *C. intybus* var. *foliosum* (Van

Stallen 2003). Les huit autres paires d'amorces étaient disponibles au laboratoire.

Tableau 3.10 : Combinaisons d'amorces pour l'étape d'amplification sélective de l'AFLP testées chez *C. intybus* et utilisées pour le génotypage de la population SAMCHIC

	M47	M56	M57	M59				
E32		*	*					
E35	*	*	*	*				
E33								
	Combinaison donnant de bons résultats chez <i>C. intybus foliosum</i> (Van Stallen 2003)							
	Combinaison donnant de bons résultats chez <i>C. intybus sativum</i> dans cette étude							
	Combinaison ne donnant pas de bon résultat chez <i>C. intybus sativum</i> dans cette étude							
*	Combinaisons utilisées pour le génotypage de la population SAMCHIC							

Les combinaisons d'amorces contenant les amorces E32 et E35 ont donné des résultats concluants par rapport au nombre de fragments amplifiés et par rapport à l'intensité du signal. En revanche, l'amorce E33 a donné des sorties AFLP dont le signal était très faible. Cette amorce n'a donc pas été retenue. Les amplifications à l'aide des paires d'amorces E32M47 et E32M59 ont, entre autres, été réalisées à l'UCL (Louvain-la-Neuve) par Claire-Noël Thienpont de l'ECAV.

Au final, six paires d'amorces (E32M56, E32M57, E35M47, E35M56, E35M57 et E35M59) ont été utilisées pour amplifier l'ADN des 200 individus de la population F2, générant ainsi 1.200 sorties AFLP. Un multiplexage des analyses a été testé, mais les résultats n'ont pas été concluants.

II.III.2. Analyse des sorties AFLP

L'étape suivante consistait à analyser chacune des sorties AFLP. Sur l'analyseur de fragments capillaire Ceq 8000, le programme d'analyse des sorties brutes contenait deux paramètres dont les valeurs ont dû être optimisées :

- La pente seuil du pic (slope threshold) : cette valeur représente l'angle que forme le pic avec l'horizontale. Plus cette valeur est faible, plus le pic est étendu et moins les valeurs sont précises. Ce critère ne possède pas d'unité.
- La hauteur seuil relative du pic en % (relative peak height threshold) : détermine la hauteur minimale pour qu'un pic soit pris en considération. Le pourcentage est calculé à partir du deuxième pic le plus intense de l'échantillon

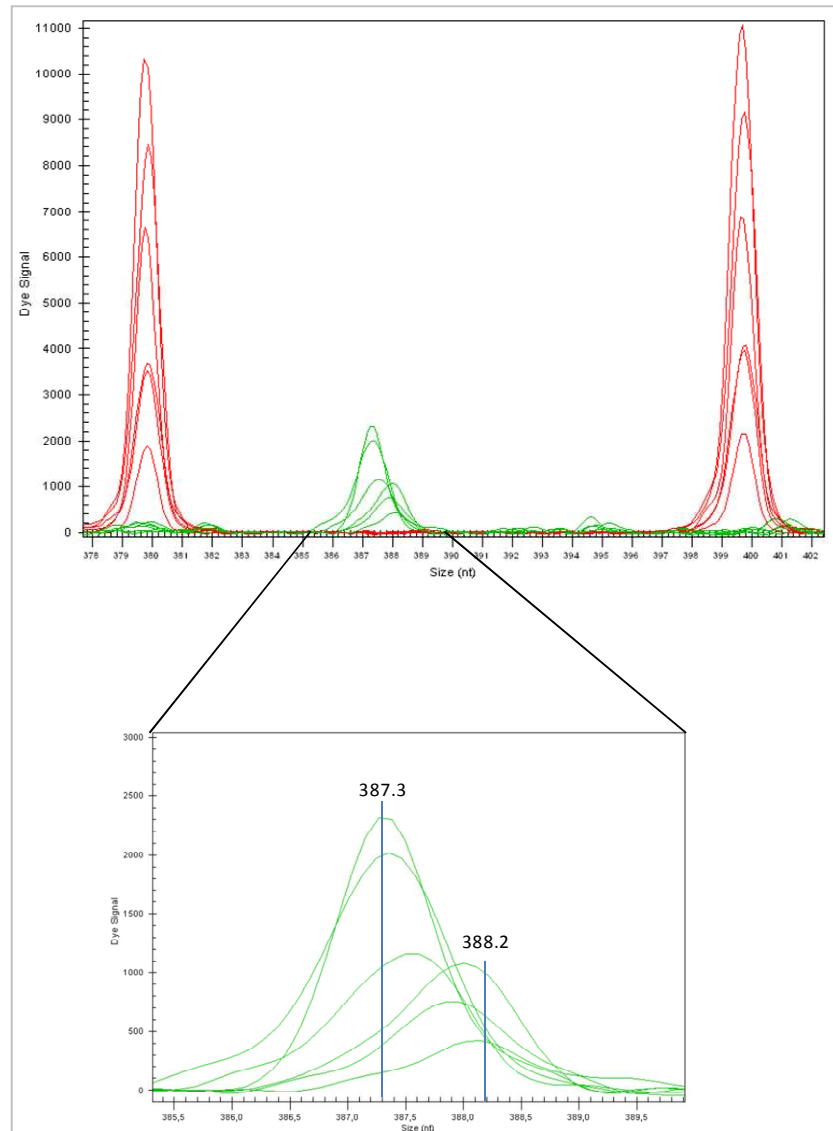


Figure 3.1 : Visualisation du phénomène d'homoplasie. Abscisse : taille des fragments en pb. Ordonnée : intensité du signal. Rouge : standard de taille. Vert : fragments obtenus par AFLP. Superposition d'une sortie AFLP de six individus. Les fragments regroupés sous un même marqueur (387) ont des tailles variant de 387,3 à 388,2

Le bruit de fond au sein des sorties variait suivant trois paramètres : l'individu qui était amplifié, la paire d'amorces utilisée pour l'amplification sélective et la taille des fragments. En effet, on a remarqué que pour les fragments de petite taille, le bruit de fond était très élevé alors qu'il diminuait pour des fragments plus grands. Cela peut s'expliquer par le fait que dans nos échantillons, le nombre de petits fragments était plus important, ce qui générait un signal plus intense et donc un bruit de fond plus conséquent. Le choix d'une valeur fixe pour les paramètres d'analyse n'était donc pas adapté à ce genre de sortie. Les valeurs de pente seuil et de hauteur relative des pics ont tout de même été définies pour chaque paire d'amorces (**tableau 3.9**) afin d'initier l'analyse des résultats. Cependant, vu la grande variabilité des 1.200 sorties, chacune d'elles a été analysée manuellement et la présence/absence de chaque pic a été examinée visuellement sur les sorties.

Lors de cette analyse, un probable phénomène d'homoplasie a également été observé : des fragments de même taille se retrouvent au sein d'un même pic alors que leur nature est différente (voir chapitre 1 point II.V.2.). Dans notre cas, cela a pu être observé via un chevauchement partiel de pics censés être de la même taille chez plusieurs individus. La **figure 3.1** illustre ce phénomène où l'on peut observer la superposition de six sorties AFLP provenant de six individus différents. Les pics regroupés sous l'étiquette « 387 » peuvent avoir une taille allant de 387,3 à 388,2 nucléotides. Dans ce cas, il n'a pas été possible de déterminer si ces pics représentaient un même marqueur ou pas. Afin d'éviter toute ambiguïté, les marqueurs AFLP présentant ce genre de profil ont été éliminés et seuls les pics montrant un chevauchement parfait ont été gardés.

L'analyse manuelle a permis de sélectionner en moyenne 40 pics par paire d'amorces pour lesquelles les profils étaient propres. Parmi ceux-ci, en moyenne sept marqueurs par paire d'amorces répondaient positivement au test χ^2 pour une distribution 3:1 ($p < 0.01$) soit 42 marqueurs au total. Après suppression des marqueurs pouvant présenter une homoplasie, seuls 13 marqueurs soit 2,6 marqueurs par paire d'amorces ont été retenus.

II.III.3. Comparaison avec les données obtenues par l'ECAV (UCL)

Claire-Noël Thienpont de l'unité d'écophysiologie et d'amélioration végétale (ECAV) de l'Université catholique de Louvain a également utilisé 12 paires d'amorces d'amplification sélective (E32M58, E32M59, E32M47, E32M61, E32M62, E39M51, E32M52, E32M50, E34M58, E32M51, E35M49, E39M58) pour le génotypage des 247 individus de la population F2 SAMCHIC. Les produits d'amplification AFLP ont migré sur gel dénaturant de polyacrylamide 6 % à l'aide d'un séquenceur Li-Cor 4200. Le nombre de marqueurs obtenus fut beaucoup plus élevé et le génotypage semblait plus évident. Au total, 163 marqueurs ont été obtenus, soit une moyenne de 13,6 marqueurs par paire d'amorces.

Le seul élément qui différait entre les deux analyses était le mode de visualisation des bandes AFLP. Dans le cas du Li-Cor, la migration conjointe des marqueurs AFLP de tous les individus analysés permettait de faire une lecture horizontale du génotype de tous les individus pour un marqueur donné. Il était donc plus évident de comparer les intensités des bandes et de déterminer dans quels cas il y avait présence ou absence du marqueur.

Dans le cas de l'analyse des sorties AFLP via le séquenceur capillaire, chaque individu a été analysé séparément et pour tous les marqueurs à la fois. Il était donc difficile de comparer les intensités d'un même marqueur chez plusieurs individus. Le bruit de fond ainsi que l'intensité du signal étant différents pour chaque sortie, le génotypage s'est avéré plus difficile que sur un séquenceur Li-Cor.

II.IV. CONCLUSION

L'AFLP a donc montré de bons résultats sur un séquenceur Li-Cor. Par contre l'analyse sur séquenceur capillaire a posé beaucoup de problèmes de génotypage. Au final, 13 marqueurs ont été développés sur séquenceur capillaire et 163 sur Li-Cor : 176 marqueurs ont donc pu être utilisés pour la cartographie. L'AFLP a dans ce cas été très utile pour obtenir un grand nombre de marqueurs. Par contre, ces marqueurs AFLP ont le désavantage d'être anonymes, difficilement transférables et dominants. Afin de compléter la carte, nous nous sommes focalisés sur des marqueurs transférables, co-dominants et

associés à des séquences transcrites tels que les microsatellites et les SNP.

III. GÉNOTYPAGE À L'AIDE DES MARQUEURS MICROSATELLITES

III.I. INTRODUCTION

Dans le chapitre 2, le polymorphisme de type microsatellite a été mis en évidence au sein des banques EST. Au final, 109 séquences ont été retenues suivant leur fonction, les caractéristiques intrinsèques du motif SSR et la possibilité de définir des amorces. Ces séquences ont été utilisées afin de développer des marqueurs microsatellites géniques pouvant être utilisés pour la cartographie.

Le présent paragraphe décrit le système utilisé pour l'évaluation préliminaire des marqueurs, les caractéristiques des SSR valides ainsi que leur utilisation pour le génotypage de la population F2 SAMCHIC.

III.II. MATÉRIEL ET MÉTHODES

III.II.1. Amplification des motifs SSR à l'aide du système à trois amorces

Le système à trois amorces (Schuelke 2000) a permis le marquage des produits PCR à l'aide d'une amorce universelle unique porteuse d'un chromophore. Les étiquettes décrites dans le **tableau 3.11** sont ajoutées aux amorces définies au sein des 109 séquences retenues dans le chapitre 2.

Tableau 3.11 : Séquence des étiquettes et de l'amorce universelle utilisés dans le système à trois amorces

Etiquette à ajouter à l'amorce FW	5'-cacgacgttgtaaacgac-3'
Etiquette à ajouter à l'amorce RV	5'-gtgtctt-3'
Amorce universelle marquée	5'-cacgacgttgtaaacgac-3'

Le mélange PCR décrit dans le **tableau 3.12** a été réalisé.

Tableau 3.12 : Mélange réalisé pour l'amplification PCR à l'aide du système « trois amorces »

	Concentration initiale	Quantité (µl)
Eau		16,9
Tampon	5 X	5
FW	1 pmol	1
ADN	50 ng/µl	0,5
RV	10 pmol	0,5
Amorce universelle marquée	10 pmol	0,5
dNTP	10 pmol	0,5
Taq	5 U/µl	0,1
	Volume total :	25

Le programme PCR utilisé est décrit dans le **tableau 3.13**.

Tableau 3.13 : Programme PCR utilisé pour le système « trois amorces »

Nombre de cycles	Durée	Température
1	5'	94°C
30	30"	94°C
	45"	56°C
	45"	72°C
8	30"	94°C
	45"	53°C
	45"	72°C
1	10'	72°C

III.III. RÉSULTATS ET DISCUSSION

III.III.1. Test d'amplification des amorces microsatellites

Un système à trois amorces a été utilisé pour l'amplification (**Figure 3.2**). Une étiquette de 19 paires de base a été ajoutée en 5' de l'amorce FW. Cette étiquette, après le premier cycle d'amplification, a été incorporée à la séquence et a servi de site d'hybridation pour une amorce universelle qui était marquée à l'aide d'un chromophore. L'avantage principal était donc qu'il ne fallait pas d'amorce spécifique marquée.

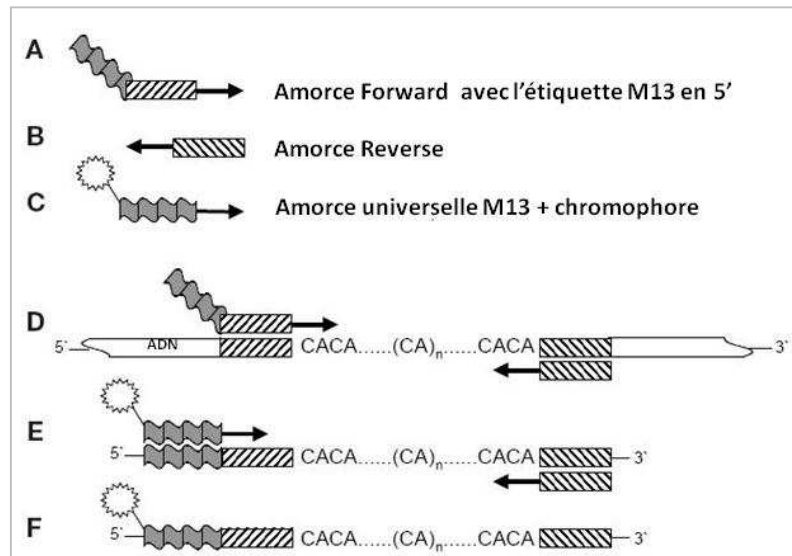


Figure 3.2 : Représentation du système d'amplification à trois amorces. Cases hachurées : séquences spécifiques du locus. Cases grises ondulées : séquence universelle M13. ◉ : Chromophore. **A, B et C :** Amorces nécessaires par locus. **D.** Lors des premiers cycles de la PCR, l'amorce Forward contenant l'étiquette M13 est incorporée au produit d'amplification **E**. Ce produit PCR est ensuite utilisé comme brin matrice pour l'amorce universelle M13 grâce à une température d'hybridation plus faible. **F.** Le produit final est marqué à l'aide d'un chromophore (Schuelke 2000)

Les paires d'amorces définies au sein des 109 séquences contenant des SSR ont été testées et 64 ont donné une amplification. Ces marqueurs SSR valides ainsi que six SSR génomiques dont nous parlerons plus bas sont décrits dans le **tableau 3.14**. Les fonctions associées à ces marqueurs via le programme Blast2Go sont également détaillées.

Afin d'évaluer le polymorphisme de ces marqueurs, une partie des paires d'amorces a été amplifiée sur huit individus appartenant à cinq variétés de chicorée différentes (Nausica, Melci, Arancha, Madonna et Vivace) (Leclerc 2007). Au final, sur 39 marqueurs testés, 27 SSR présentaient du polymorphisme, soit 70%. Le nombre moyen d'allèles pour ces marqueurs était de trois avec un maximum de cinq allèles. Les douze autres marqueurs ne montraient pas de polymorphisme chez les huit individus analysés. Cela ne signifie pas qu'ils ne sont pas polymorphes mais seulement que l'on n'a pas pu l'observer dans l'échantillon utilisé.

Tableau3.14 : Récapitulatif des 70 marqueurs microsatellites valides. Gris : SSR ayant une distribution mendélienne 2 : 1 : 1 dans la population F2

Nom du marqueur	Type de SSR	Longeur du SSR	Motif SSR	Amorce Forward	Amorce Reverse	T° d'hybridation des amorces	Estimation du polymorphisme	Nom de la séquence associée au marqueur	Fonction associée par homologie à la séquence	valeur du test Blast Hit	Processus lié à la fonction
es11	trinucleotide	21	ATC	tcacgaaccactaaaacaca	ctgaagcatgacacagcaga	56°C	non testé	REL008C07_GB	SKU5; copper i on binding	2.53E-62	Croissance
es14	trinucleotide	19	AAC	gaagatctaccaccaatcc	ataaggctgaagaatgaag	56°C	4 allèles chez 8 individus	RKL005E04_XT	putative beta (1,2)-xylosyltransferase	1.39E-37	métabolisme des carbohydrates
es15	trinucleotide	18	AAT	tacatcatctgtccaccc	taatacgtgtacagccaac	56°C	4 allèles chez 8 individus	NG010G12_XT	beta-fructofuranosidase acid invertase	1.24E-33	métabolisme des carbohydrates
es16	trinucleotide	14	AT	aaatgaagctgtgtgtgtgt	ctgtgcaatctcaaac	56°C	3 allèles chez 8 individus	C2CS252.11_C24.ab1	gdp-mannose pyrophosphorylase	1.00E-24	métabolisme des carbohydrates
es17	trinucleotide	30	ATC	acaaagatgtgtctctcaaaa	acgtgtgacgtctctaaag	56°C	4 allèles chez 8 individus	RBL009C06_GB	SubSNF1 protein	1.41E-01	métabolisme des carbohydrates
es18	trinucleotide	21	ATC	gggtcaggtctgtgaagctc	aaacaaacccctgttaaac	56°C	4 allèles chez 8 individus	CCIM658.b.1_P05.ab1	trehalase-6-phosphate synthase	1.00E-60	métabolisme des carbohydrates
es19	trinucleotide	33	CTT	agacatcatctgttcaag	ggaatagggagacatttc	56°C	3 allèles chez 8 individus	KC12_Conf.g5.768	glycoside hydrolase family 28 protein polygalacturonasemifin protein	1.00E-57	métabolisme des carbohydrates
es110	trinucleotide	15	TGC	gtctgtgtgtgtgtgtgtgt	ctctgtgtgtgtgtgtgtgt	56°C	3 allèles chez 8 individus	KC12_Conf.g4.483	glycosyl hydrolase family 20 protein	1.00E-81	métabolisme des carbohydrates
es111	trinucleotide	16	AT	ctctacagcacaacacag	tgagagcaggtgtgtgaagc	56°C	4 allèles chez 8 individus	REN002H02_SC_F	NUDIX hydrolase; Glycosyl transferase, family8	7.33E-41	métabolisme des carbohydrates
es112	trinucleotide	15	AAT	dgctctcaaggaagaatgtc	cccaacatgtcaattctagg	56°C	3 allèles chez 8 individus	Contig7407	putative A gonauie protein	1.00E-118	développement des racines
es113	trinucleotide	18	AAC	agtggtgtgttccctccaacc	cttggagggtgtgtataaaga	56°C	3 allèles chez 8 individus	Contig8277	Cs10g032490.0	1.00E-14	développement des racines
es114	trinucleotide	12	AG	gctctctacacgtcaattc	ctctgtgtgtgtgtgtgtgt	56°C	2 allèles chez 8 individus	Cich_iny.EH682296	br1 like 1 protein-like	1.00E-40	transport
es115	trinucleotide	14	CCG	gctcttctgtctctctctc	gtgtgtgtgtgtctctctctc	56°C	2 allèles chez 8 individus	Cich_iny.EH691306	hsp1a-exonin 3	1.00E-91	organisation et biogenese des parois végétales
es116	trinucleotide	15	AAG	gagcagagaagaataatgttg	taataacacgggagctgtc	56°C	1 allèles chez 8 individus	Contig7838	unknown	1.00E-33	croissance
es117	trinucleotide	18	AAG	cagacacgtgaaaagaatccc	aaacacaaaatcgtctagg	56°C	2 allèles chez 8 individus	Cich_iny.EH706274	lin g-h2 finger protein at1q	1.00E-72	réponse à la chéline
es118	trinucleotide	18	ACT	ctgtgaaggtgtgtgataaagc	gttaagacatctatctatgc	56°C	1 allèles chez 8 individus	Cich_iny.EH705156	LEA14_G05HILate embryogenesis abundant protein Leat1-4-A	1.00E-53	organisation et biogenese des parois végétales
es119	trinucleotide	26	AG	agaaagaaagtggtgtgtgtgtgt	gtgtgtgtgtgtgtgtgtgtgt	56°C	4 allèles chez 8 individus	Contig130	DC1 2 homologue	1.00E-49	modification posttraductionnelle des proteines
es120	trinucleotide	15	AGG	atgtgtgtgtgtgtgtgtgtgt	ctttgtgtgtgtgtgtgtgtgt	56°C	2 allèles chez 8 individus	Contig6233	hypothetical protein Os1_015624	1.00E-57	organisation et biogenese des parois végétales
es121	trinucleotide	18	AGG	cagtgagacaacaacacacac	tgatactcgtgtgtgtgtgt	56°C	2 allèles chez 8 individus	Cich_iny.EH680832	end o-1-3-beta-glucanase-like protein	1.00E-67	métabolisme des carbohydrates
es126	trinucleotide	15	CCG	gccaacttctctatacagc	tcagtttccgtatgtgtgaagc	56°C	non testé	Contig2166	lutein-induced protein	1.00E-71	développement des racines
es127	trinucleotide	24	AAC	tttgaggtgtgtgtgtgtgtgtgt	atggaagaacatgaataatgt	56°C	2 allèles chez 8 individus	RKL007C05_GB	VT521-8-like protein	1.00E-52	métabolisme des carbohydrates
es128	trinucleotide	14	AC	gtgtgttcaagagatgac	ctgtgttctcaactctgt	56°C	4 allèles chez 8 individus	C2CS14242.1_C28.ab1	cytosolic phosphate-dependent phosphofructo-1-kinase	1.00E-49	métabolisme des carbohydrates
es130	trinucleotide	36	ATC	gataaagggcctccattc	ctcgtctctagactctgtc	56°C	non testé	Cich_iny.EH688076	transferase, transferring glycosyl group / transferase, transferring hexosyl groups	1.00E-20	modification posttraductionnelle des proteines
es132	trinucleotide	12	AG	actgaagcagaataaaccttc	ctctgtgtgtgtgtgtgtgt	53.5°C	3 allèles chez 8 individus	Contig4614	Protein of unknown function DU0707	1.00E-74	saison aux sucres
es133	trinucleotide	16	AC	ctctgtgtgtgtgtgtgtgtgt	gtctgtgtgtgtgtgtgtgtgt	56°C	non testé	Contig2106	DCL_LVESCSDCL protein, chloroplast precursor	1.00E-64	modification posttraductionnelle des proteines
es134	trinucleotide	18	ATC	taaatgttctcagaagaataatgt	ctctgtgtgtgtgtgtgtgtgt	56°C	4 allèles chez 8 individus	Contig6953	HP74021, hypothetical protein	1.00E-12	métabolisme des carbohydrates
es138	trinucleotide	12	AT	ctctgtgtgtgtgtgtgtgtgt	tcctaaagattctctctgt	56°C	3 allèles chez 8 individus	KC12_Conf.1279	phosphofructokinase family protein	1.00E-59	métabolisme des carbohydrates
es140	trinucleotide	15	CGT	tctctgtgtgtgtgtgtgtgtgt	gttaaagctatctatctctc	56°C	non testé	KC12_Conf.g5886	lactoyglutathione lyase family protein glyoxalase I family protein	1.00E-66	métabolisme des carbohydrates
es141	trinucleotide	24	AAG	gagagtgatgagtgatgagtg	gtgtgtgtgtgtgtgtgtgtgt	56°C	non testé	Contig2365	unknown	1.00E-307	regulation de la transcription
es143	trinucleotide	21	AAC	gagatgaatgagatgataaac	gggaatgacttctatctc	56°C	non testé	Contig785	AF488561, putative NHLH transcrition factor	1.00E-11	regulation de la transcription
es146	trinucleotide	21	AAG	dgctctcaatgtgtgtgtgtgt	atctgtgtgtgtgtgtgtgtgt	56°C	3 allèles chez 8 individus	Contig5901	ap2 erf dom ain-containing transcription factor	1.00E-60	regulation de la transcription
es148	trinucleotide	20	AG	cccatctcaggagaagatctagg	ctaacaggccatctacatttc	53.5°C	non testé	Contig8874	ncular H+-ATPase C subunit	1.00E-59	transport
es149	trinucleotide	21	AAG	cgctgtgtgtgtgtgtgtgtgt	cgatatctgtgtgtgtgtgtgt	56°C	1 allèles chez 8 individus	Contig8723	AF196289_15 erfThr specific protein phosphatase 2AB regulatory subunit beta isoform	1.00E-73	transduction du signal
es150	trinucleotide	21	ATC	gggtgtgtgtgtgtgtgtgtgt	ctacttctctaacactctca	56°C	non testé	Contig4125	lact tolerance protein	1.00E-118	réponse à la déshication
es151	trinucleotide	36	AG	cccacacaaattctctctc	ctattctctgtatggcgtgt	56°C	non testé	Contig326	ubiquitin-conjugating enzyme 1	1.00E-79	regulation de la transcription
es153	trinucleotide	42	AG	cttctctctctctctctctc	tcagcttctctctctctgt	56°C	3 allèles chez 8 individus	Contig3267	AF510215_1n am-like protein 18	1.00E-85	regulation de la transcription
es156	trinucleotide	33	ACC	tgggtgtgtgtgtgtgtgtgtgt	agaaggtgtgtgtgtgtgtgt	53.5°C	non testé	Cich_iny.EH710670	BONZAH	1.00E-51	réponse de defense
es157	trinucleotide	32	AG	tcatctgtgtgtgtgtgtgtgt	atcaccagtaacatgtatcac	56°C	non testé	Cich_iny.EH685237	unknown protein	1.00E-89	Croissance
es159	trinucleotide	30	AG	agcagatgtgtgtgtgtgtgtgt	ttccaccacagctatctc	53.5°C	non testé	Contig4555	gfp-binding rab2a like pro lein	1.00E-101	transduction du signal
es160	trinucleotide	27	ATC	ggctgtgtgtgtgtgtgtgtgt	agacacatcatctcaacttc	56°C	non testé	Cich_iny.EH708379	hypothetical protein	1.00E-13	regulation de la transcription
es161	trinucleotide	24	AAT	algtttctctctctctctctc	gtctattaagagacaagaagc	53.5°C	non testé	Contig7954	CCAAT-binding factor B sub unit homolog	1.00E-17	regulation de la transcription
es162	trinucleotide	22	AG	tgaccagagagagatgtatgag	gaagcagatattctctagacag	56°C	3 allèles chez 8 individus	Contig5073	NSK6, Shaggy-like kinase 6	1.00E-61	modification posttraductionnelle des proteines
es164	trinucleotide	21	AAG	agagagatgtgtgtgtgtgtgt	agatctccgtgaattatgagc	56°C	3 allèles chez 8 individus	Cich_iny.EH677574	Chd1p0511100	1.00E-28	-
es165	trinucleotide	21	AAT	gtgtgtgtgtgtgtgtgtgtgtgt	ttatgtgtgtgtgtgtgtgtgt	56°C	non testé	Contig1881	ATP binding / isoleucine-RN A ligase / RNA ligase	1.00E-113	reduction
es166	trinucleotide	21	ATC	gtgtgtgtgtgtgtgtgtgtgtgt	atcagaagcgtatattctc	53.5°C	1 allèles chez 8 individus	Contig2779	heme oagc enase 2	1.00E-35	photosynthese
es167	trinucleotide	21	AAC	ctcaagagagcgaactctatgc	taattacacgctcttaccagc	56°C	3 allèles chez 8 individus	Contig4061	unknown pro tem	1.00E-68	-
es168	trinucleotide	20	AG	gtatgtgtctgtgtgtgtgtgtgt	atggagactagattctctcc	56°C	non testé	Contig3683	senesence-associated putat ve protein, putative	1.00E-81	modification post-traductionnelle des proteines
es169	trinucleotide	20	AG	tactgtgtgtgtgtgtgtgtgtgt	cggttatctgtgaattatgtgt	56°C	3 allèles chez 8 individus	Contig5565	PBS1 (AVRPP HB SUSCEPTIBLE 1), kinase	1.00E-100	réponse de defense
es171	trinucleotide	40	AG	gtgtgtgtgtgtgtgtgtgtgtgt	gtgtgtgtgtgtgtgtgtgtgtgt	56°C	non testé	Cich_iny.EH679186	Protease-associated PA. Peptidase A2/B, minor histocompatibility antigen H13	1.00E-70	modification post-traductionnelle des proteines
es173	trinucleotide	18	AC	atcacaaggtgtgtgtgtgtgt	ctactgtgtgtgtgtgtgtgtgt	56°C	3 allèles chez 8 individus	Cich_iny.EH708630	putative RING zinc finger protein	1.00E-44	réponse à la chéline
es174	trinucleotide	16	AG	actctatgtctcaacaagtg	cagctctatgtcttaattgt	56°C	non testé	Contig3375	peptidase / proglutamy- pept idase I	1.00E-78	modification posttraductionnelle des proteines
es175	trinucleotide	16	AG	atctccagaaagtcacagatcc	tcgtctccatttctcaagc	53.5°C	1 allèles chez 8 individus	Contig8426	Ctr zipper r transport	1.00E-45	transport
es180	trinucleotide	18	ACC	agctccagatgtgtgtgtgtgt	gtgtgtctatctgtatgacc	56°C	1 allèles chez 8 individus	Contig3818	protein tra nsport SEC13-like protein	1.00E-106	transport
es182	trinucleotide	18	AAG	cagatgtgtgtgtgtgtgtgtgt	ctgtatctgtgtgtgtgtgtgt	53.5°C	1 allèles chez 8 individus	Contig4891	late-ato ndant protein	1.00E-147	modification post-traductionnelle des proteines
es183	trinucleotide	18	AAT	gagacatcatattctatcc	ggagatgtgtataactatgt	53.5°C	non testé	Contig2186	Cytochrome vacuolar atpase, subunit h protein	1.00E-25	transport
es184	trinucleotide	18	AAG	agacccaacaattacagacac	gtgtatgtgtgtgtgtgtgtgt	56°C	1 allèles chez 8 individus	RGL006_F08	beta-tubul in 1	1.00E-19	transport
es198	trinucleotide	18	ATC	cttctgtctcttctctctc	ctaacacagagatctacaaag	56°C	3 allèles chez 8 individus	Contig83	hypothetical ring-h2 zinc finger protein	1.00E-13	raison / Non zinc
es199	trinucleotide	18	ACT	gggtgtgtgtgtgtgtgtgtgt	cttctcttcaacagaagacaa	56°C	2 allèles chez 8 individus	Contig761	hydrogen-tra nsporting ATP synthase	1.00E-52	transport
MS1	trinucleotide	36	TGA	ctaacacagagctctcaaaagac	ggatgtgtgtgtgtgtgtgtgtgt	56°C	non testé	RBL007C09	serine palmitoyltransferase	9.04E-73	modification post-traductionnelle des proteines
MS3	trinucleotide	21	AAT	ctctgtctctctctctc	ctaacacagagatctcaacag	56°C	non testé	RGL004_O10	putative pyruvate dehydrogenase kinase, S' partial	1.00E-22	métabolisme des carbohydrates
MS5	trinucleotide	24	TGA	gggtgtgtgtgtgtgtgtgtgt	cttctctctcaacaagaagaa	56°C	non testé	RGL008H04	nucleosome chromatin a ssembly factor group A	7.00E-09	-
MS10	trinucleotide	20	CTG	gagcagatcatattctaccc	ggaggtgtgtgtgtgtgtgtgt	56°C	non testé	RBL005F06	expressed protein	7.84E-44	-
MS11	trinucleotide	18	AG	agacccaacaattacagacac	gtgtatgtgtgtgtgtgtgtgt	56°C	non testé	RGN040B1	serine palmitoyltransferase	1.04E-12	Apoptose
MS12	trinucleotide	18	AC	agagatgtgtgtgtgtgtgtgt	gtgtgtgtgtgtgtgtgtgtgt	56°C	non testé	RKL004C06	unknown protein	1.00E-64	-
cg16	trinucleotide	24	GA	ttccacacagctgtctacatctc	cttctctgtgtgtgtgtgtgtgt	58.1	7 allèles chez 8 individus	cg16	non codant	-	-
cg163	trinucleotide	22	GA	gaagagaatgtgtgtgtgtgt	ttcttctgtgtgtgtgtgtgtgt	58.7	17 allèles chez 54 individus	cg163	non codant	-	-
cg164	trinucleotide	32	GA	gagcctctgtgtgtgtgtgtgt	gttcttctgtgtgtgtgtgtgtgt	58.5	3 allèles chez 8 individus	cg164	non codant	-	-
cg165	trinucleotide	32	GA	ccatctacagaagaagcattt	tttcttctctctgtgtgtgtgtgt	58.7	14 allèles chez 54 individus	cg165	non codant	-	-
cg166	trinucleotide	22	GA	ttctgtgtgtgtgtgtgtgtgt	tttcttgaagaatgtgtgtgtgt	58.7	10 allèles chez 54 individus	cg166	non codant	-	-
cg172	trinucleotide	24	GA	ttctgtgtgtgtgtgtgtgtgt	gttcttgaagaatgtgtgtgtgt	57.8	6 allèles chez 6 individus	cg172	non codant	-	-

Parmi les marqueurs SSR testés à l'aide du système à trois amorces, six marqueurs supplémentaires (marqueurs Ciga dans le **tableau 3.14**) étaient des microsatellites provenant d'une banque d'ADNg enrichie réalisée en 2004 au laboratoire (Muys 2004). Ces marqueurs sont considérés comme non-géniques puisqu'à ce jour, aucune homologie n'a été mise en évidence avec des séquences codantes. Le polymorphisme de ces marqueurs a également été testé. Le nombre d'allèles détectés varie entre 17 et 3. On a donc pu observer chez ces marqueurs une plus grande variabilité par rapport aux EST-SSR. Cela peut s'expliquer d'une part par le nombre plus élevé d'individus analysés, mais également par le fait que ces microsatellites se situent probablement dans des régions non codantes de l'ADN. La pression de sélection étant moins forte dans ces régions, les SSR y évoluent plus vite et présentent plus de polymorphisme.

Au total, 70 marqueurs SSR ont été validés.

III.III.2. Génotypage des 200 individus du croisement SAMCHIC

Afin de pouvoir être cartographiés, les marqueurs devaient présenter du polymorphisme au sein du croisement SAMCHIC. Chaque marqueur a donc été utilisé pour amplifier l'ADN de l'individu provenant de la F1 et choisi pour générer la F2. Seuls les marqueurs hétérozygotes ont été utilisés pour génotyper les 200 individus de la F2. La **figure 3.3** illustre le cas d'un marqueur hétérozygote (MS10) et d'un marqueur homozygote (MS5) amplifiés sur l'individu F1. Au total, 17 marqueurs SSR sur les 70, soit 24% présentaient du polymorphisme au sein du croisement.

Ce pourcentage relativement faible peut s'expliquer par la localisation des SSR dans les régions transcrites du génome moins polymorphes par rapport aux régions non transcrites. Chez *C. intybus*, les SSR génomiques utilisés pour la construction de trois cartes génétiques (Cadalen et al. 2010) présentaient respectivement 51 %, 41,6 % et 26,7 % de polymorphisme dans les populations de ségrégation intraspécifiques. D'autre part, des marqueurs SSR géniques et génomiques ont été utilisés pour génotyper deux populations de ségrégation intraspécifiques de *Brassica rapa* (Ramchiary et al. 2011). 41,7 % et 67 % des SSR génomiques présentaient du polymorphisme dans ces deux populations contre 14 % et 20 % pour les SSR géniques.

Le faible pourcentage de marqueurs SSR polymorphes obtenu dans ce travail pourrait donc être expliqué par la nature de ces marqueurs ainsi que par une faible diversité génétique au sein de la population SAMCHIC.

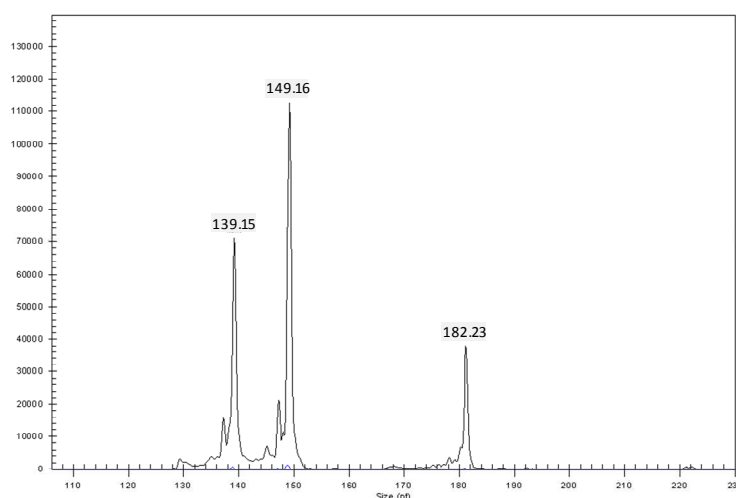


Figure 3.3 : Produit d'amplification de l'individu F1 du croisement SAMCHIC obtenu à l'aide des marqueurs SSR MS5 (182 pb) et MS10 (139 et 149 pb). Abscisse : taille des fragments en pb. Ordonnée : intensité du signal

Les 17 marqueurs ont été amplifiés sur les 200 individus de la F2. Afin de diminuer le nombre d'analyses, plusieurs marqueurs ont été amplifiés et analysés ensemble. Ce multiplexage était possible entre marqueurs de tailles et de chromophores différents (**figure 3.4**).

Pour vérifier que les marqueurs SSR avaient bien une distribution 2:1:1 au sein de la F2, un test χ^2 ($p < 0.01$) a été réalisé. Celui-ci a montré une distorsion de ségrégation pour trois marqueurs. Ceux-ci ont donc été supprimés. Au final, 14 marqueurs microsatellites (12 EST-SSR et deux SSR génomiques) ont pu être cartographiés (en gris dans le **tableau 3.14**). La société Chicoline S.A. a également génotypé la population F2 à l'aide de 27 microsatellites non-géniques obtenus via une banque enrichie. Ces marqueurs ont également été cartographiés.

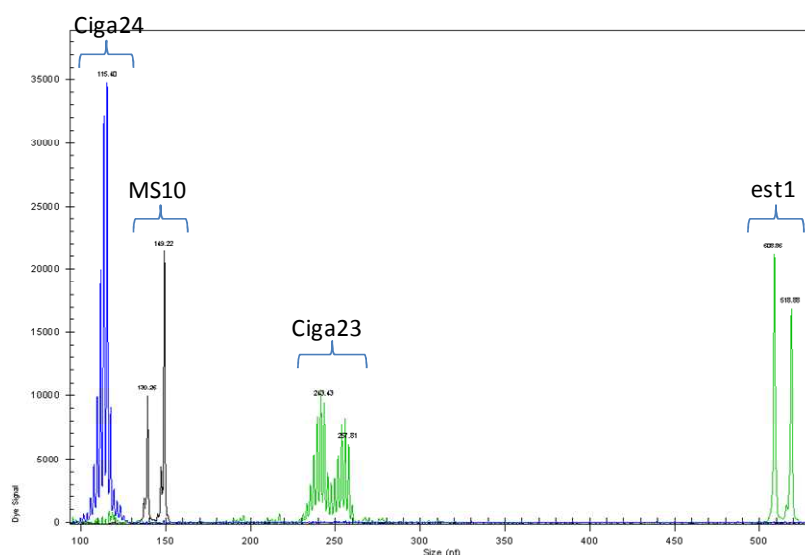


Figure 3.4 : Exemple de multiplexage des produits d'amplification d'un individu de la population F2 obtenu à l'aide de 4 paires d'amorces SSR. Abscisse : taille des fragments en pb. Ordonnée : intensité du signal

III.IV. CONCLUSION

Au total, dans cette analyse, 70 marqueurs SSR comportant 64 marqueurs géniques et 6 marqueurs génomiques ont été développés et mis au point. Parmi les marqueurs SSR utilisés pour génotyper huit individus provenant de cinq cultivars différents, 70 % présentaient du polymorphisme. Le pourcentage de marqueurs polymorphes au sein de la population SAMCHIC fut cependant beaucoup plus faible puisque sur 70 marqueurs, seuls 17 soit 24 % ont été retenus. Après analyse de la distribution des génotypes, trois marqueurs ont été écartés car ils présentaient une distorsion de ségrégation. Au final, 14 marqueurs SSR (12 géniques et 2 génomiques) ont pu être utilisés pour la cartographie.

IV. GÉNOTYPAGE À L'AIDE DES MARQUEURS SNP

IV.1. INTRODUCTION

Vu le faible nombre de marqueurs SSR polymorphes dans la population SAMCHIC, nous nous sommes focalisés sur le développement de marqueurs SNP. Ces marqueurs présentent de manière générale moins de polymorphisme que les marqueurs SSR ; cependant, ils sont plus nombreux dans le génome. En effet, dans le chapitre précédent, plus de 584 SNP ont été mis en évidence au sein de 158 séquences transcrites (**tableau 2.7**, chapitre 2).

Plusieurs méthodes de génotypage ont été testées et utilisées pour l'analyse des SNP. Une méthode d'extension-ligation ainsi qu'une méthode de ligation ont été développées afin de pouvoir faire du génotypage à haut débit sur microarray. Ensuite, vu le nombre plus faible qu'initialement prévu de SNP à analyser dans la population SAMCHIC, deux méthodes basées sur l'extension d'amorces (PAMSA et Bi-PASA) ont été utilisées afin de visualiser les allèles sur séquenceur capillaire et sur gel d'agarose.

Le présent chapitre décrit la mise au point des méthodes de génotypage ainsi que leur utilisation pour l'analyse du croisement SAMCHIC.

IV.II. MATÉRIEL ET MÉTHODES

IV.II.1. Méthode d'extension-ligation

Le mélange ainsi que le programme d'extension-ligation décrits dans le **tableau 3.15** ont été réalisés.

Tableau 3.15 : Mélange et programme d'extension-ligation

	Quantité (µl)			
Produit PCR	0,25			
Taq DNA ligase (40U/µl)	0,25			
Taq DNA polymerase (5U/µl) (NEB)	0,2			
LSO (10µm)	0,5			
ASO-T7 (10 µm)	0,5			
ASO-T3 (10 µm)	0,5	1 x	2 min	95°C
Tampon polymerase 5x	5	40 x	30sec	94°C
Volume total	25		30 sec	45°C

Tableau 3.16 : Mélange et programme utilisés pour l'amplification universelle des produits d'extension-ligation

	Quantité (µl)			
Produit d'extension-ligation	1 (dilué 100 x)			
dNTP (10 µM)	0,5			
T3 (10 µM)	0,5			
T7 (10 µM)	0,5	1 x	2 min	94°C
FW (10 µM)	0,5	25 x	25 sec	94°C
Taq Polymerase (5 U/µl)	0,2		35 sec	55°C
Tampon (5 x)	5		15 sec	72°C
Eau	15,3			
Volume total	25	1 x	2 min	72°C

Les produits obtenus ont ensuite été précipités et lavés à l'éthanol. Ceux-ci ont été utilisés pour réaliser une amplification à l'aide d'amorces universelles T3 et T7. Le mélange et le programme décrits dans le **tableau 3.16** ont été effectués.

Les fragments étant marqués par un chromophore grâce aux amorces T3 et T7 lors de la PCR universelle, ceux-ci ont pu être visualisés au séquenceur capillaire.

IV.II.2. Méthode de ligation

Le mélange ainsi que le programme de ligation décrits dans le **tableau 3.17** ont été réalisés.

Tableau 3.17 : Mélange et programme utilisés pour la ligation

	quantité (µl)			
Produit PCR	0,25			
Ligase (40U/µl)	0,2			
LSO (10µm)	0,25			
ASO-T7 (10 µm)	0,25			
ASO-T3 (10 µm)	0,25	1x	5 min	95°C
Tampon ligase (10x)	1	3 x	30 sec	95°C
Volume total	10 µl		20 min	45°C

Les produits obtenus ont ensuite été précipités et lavés à l'éthanol. Ceux-ci ont été utilisés pour réaliser l'amplification à l'aide des amorces universelles T3 et T7 (point IV.II.1). Les fragments ont ensuite été analysés au séquenceur capillaire.

IV.II.3. Méthode PAMSA

Deux amorces spécifiques d'allèles et une amorce spécifique de locus ont été définies à partir des critères suivants :

- Tm autour de 50°C
- En partant de l'extrémité 3', une mutation (A/T→G/C) au niveau de la 4^{ème} base propre à chaque amorce spécifique d'allèle a été ajoutée
- Une étiquette de 5 bases sur l'amorce spécifique de l'allèle 1 et une de 15 bases pour l'amorce spécifique de l'allèle 2 ont été ajoutées

Chacun des trois oligos en solution stock de 100 µM a été dilué 30 fois dans le même mélange.

Le mélange ainsi que le programme PCR décrits dans le **tableau 3.18** ont été réalisés.

Tableau 3.18 : Mélange et programme PCR réalisés pour la méthode PAMSA

	Quantité (µl)			
Tampon 10 x	1			
Mix 3 amorces (3,3 µM)	0,3			
dATP 10 mM	0,1			
dCTP 10 mM	0,1			
dGTP 10 mM	0,1			
dTTP 10 mM	0,08	Nombre de cycles	Température	Temps
dUTP marqués 1 mM	0,08	1 x	94°C	4'
Taq polymérase(5U/µl)	0,06	35 x	94°C	20"
ADN (10 ng/µl)	0,5		50°C	45"
Eau	7,68		72°C	30"
Volume total	10	1 x	72°C	7'

Les dUTP associés à un chromophore ont été incorporés au produit d'amplification lors de la PCR, ce qui a permis d'obtenir des fragments marqués sans utiliser des amorces couplées à un chromophore.

IV.II.4. Méthode Bi-PASA

Deux amorces spécifiques d'allèles et deux amorces spécifiques de locus ont été définies à partir des critères suivants :

- Le T_m des amorces P et Q (spécifiques de locus) devait être inférieur de 20 à 25°C par rapport au T_m du fragment PQ (idéalement compris en 300 et 1000 pb)
- Le T_m des amorces A et B (spécifiques d'allèles) devait être inférieur d'environ 35°C par rapport au fragment PQ
- La température d'hybridation du programme PCR devait être de 20°C inférieure au T_m du fragment PQ.
- Une étiquette GC d'environ 10 bases (GGGGCGGGGC) a été ajoutée en 5' des sondes A et B

Le mélange ainsi que le programme PCR décrits dans le **tableau 3.19** ont été réalisés.

Tableau 3.19 : Mélange et programme PCR réalisés pour la méthode Bi-PASA

	Quantité (µl)			
Eau	17,55			
Tampon 5 x	5			
dNTP (10 mM)	0,25			
A (10 µM)	0,25	Nombre de cycles	Température	Temps
B (10 µM)	0,25	1x	95°C	2 min
P (10 µM)	0,25	30x	95°C	15 sec
Q (10 µM)	0,25		T° d'hybridation	30 sec
Taq polymérase(5 U/µl)	0,2		descendre à 72°C en 1 min	
ADN (100 ng/µl) dilué 10 x	1		72°C	2 min
Volume total	25	1x	72°C	10 min

Les produits PCR ont ensuite été déposés sur un gel d'Agarose 1% afin de visualiser la taille des différents fragments.

IV.III. RÉSULTATS ET DISCUSSION

IV.III.1. Mise au point des méthodes de génotypage

IV.III.1.1. Méthode d'extension-ligation

Vu le nombre important d'individus à génotyper ainsi que le nombre de marqueurs potentiels, une méthode à haut débit semblait nécessaire. Dans un premier temps, le système GoldenGate développé par Illumina semblait un choix approprié.

La méthode GoldenGate (**figure 3.5**) repose sur la conception de trois sondes pour chaque locus :

- deux sondes spécifiques d'allèles (ASO, Allele Specific Oligonucleotide) dont la dernière base en 3' correspond au SNP associées à une séquence d'amorces universelles (P1 ou P2)
- une sonde spécifique de locus (LSO, Locus specific Oligonucleotide) dont le site d'hybridation est situé à quelques bases du SNP. Cette sonde est également composée d'une séquence d'adressage ainsi que de la séquence d'une amorce universelle P3.

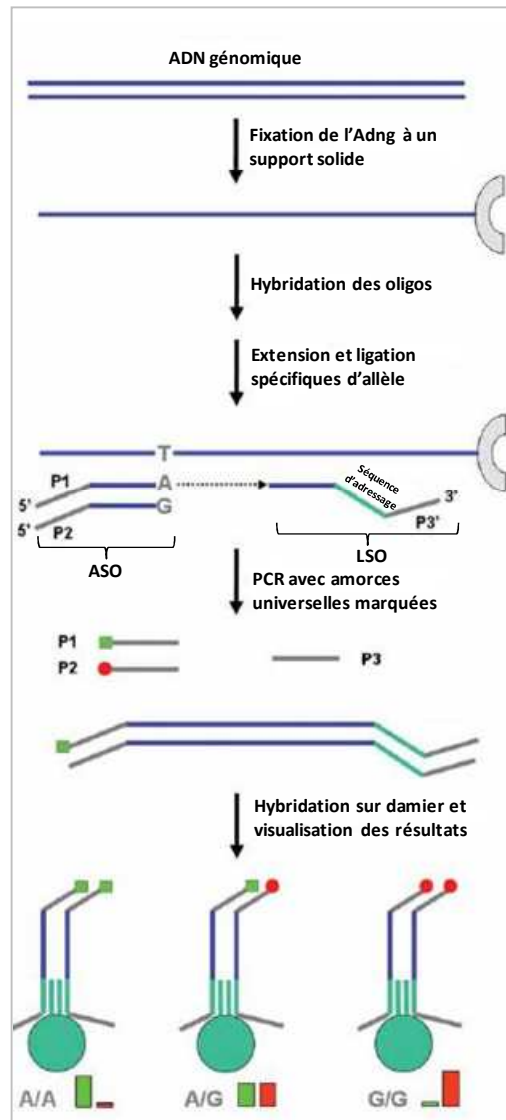


Figure 3.5 : Schéma représentant le système GoldenGate pour le génotypage des SNP (Illumina). ASO : oligo spécifique d'allèle. LSO : oligo spécifique de locus. P1, P2, P3 : amorces universelles. ■ et ● : Fluorochrome. L'ADNg est fixé sur un support solide. L'hybridation d'un ASO et d'un LSO sur l'ADN cible permet de réaliser une extension et une ligation des deux sondes. Le fragment obtenu sert de matrice lors de la PCR à l'aide des amorces P1, P2 (suivant l'allèle) et P3. La séquence d'adressage spécifique du locus permet l'hybridation des fragments sur un damier. La fluorescence émise pour chaque emplacement du damier informe sur la nature du ou des allèles du locus d'intérêt (Fan, 2003)

Lors de la première étape, l'ADN fragmenté est associé à un support solide. Cela permet de fixer l'ADN et donc ensuite de supprimer par lavage, à chaque étape, les sondes qui n'ont pas été hybridées et cela sans éliminer l'ADN matrice.

La deuxième étape consiste à mettre en présence de l'ADN matrice les trois sondes décrites plus haut. Une polymérase ainsi qu'une ligase sont ajoutées. Dans le cas de l'hybridation d'un ASO et d'un LSO sur un même brin, la polymérase vient allonger l'ASO en 3' et une fois l'écart comblé, la ligase fusionne les deux sondes ensemble.

L'étape suivante consiste à réaliser une PCR universelle à l'aide des amorces P1 et P2 marquées avec des fluorochromes de couleurs différentes ainsi qu'avec l'amorce P3. Les fragments ainsi générés peuvent être hybridés sur un damier comportant des billes associées à des séquences de capture complémentaires aux séquences d'adressage présentes dans la sonde LSO.

La dernière étape comprend la lecture de la fluorescence de chaque damier. La couleur ainsi détectée permet de définir s'il s'agit d'un échantillon homozygote AA ou BB ou encore d'un hétérozygote AB.

Les coûts associés à la mise en place de ce type de plateforme ne pouvaient être pris en charge ni par notre laboratoire ni par les sélectionneurs. Il a donc été décidé de développer un système de génomique similaire, à moindre coût et à plus petite échelle. Afin de garder les avantages de la méthode GoldenGate, les caractéristiques suivantes ont été maintenues :

- L'extension-ligation des sondes : le couplage des deux techniques permet d'augmenter la sensibilité et donc la spécificité de la détection des allèles (Abravaya et al. 1995). En effet, la sonde ASO doit être parfaitement hybridée et donc parfaitement spécifique pour qu'il y ait extension par la polymérase et ligation par la ligase.
- Les amorces associées à un fluorochrome sont universelles. Cela permet donc d'éviter de marquer des sondes spécifiques de chaque locus. Le coût est donc fortement diminué.
- L'utilisation de séquences d'adressage permet d'identifier facilement l'origine des fragments. Ces séquences peuvent être utilisées pour l'hybridation sur damier, mais elles peuvent

également être utiles si l'on veut analyser ces fragments par séquençage.

- La méthode GoldenGate ne demande pas d'amplification au préalable des loci. Cela permet donc de supprimer une étape. Il a fallu cependant vérifier que dans notre cas, cette étape n'était pas nécessaire.

La méthode dite d'extension-ligation (**figure 3.6**) que nous avons mise au point est basée sur différents articles (Abravaya et al. 1995; Fan et al. 2003; Macdonald et al. 2005).

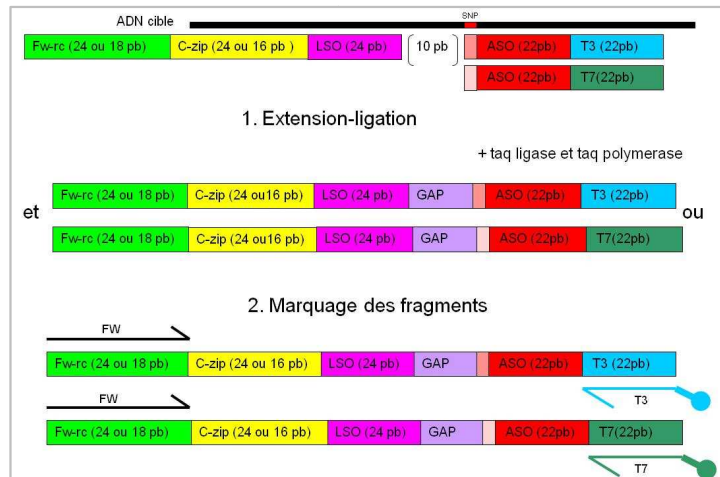


Figure 3.6 : Méthode d'extension-ligation. T3, T7 et FW : amorces universelles. ASO : oligo spécifique d'allèle. LSO : oligo spécifique de locus. C-Zip : séquence d'adressage. Fw-rc : « reverse complement » de l'amorce FW. et : chromophore. : emplacement du SNP. Les sondes spécifiques d'allèle, une fois parfaitement hybridées sur l'ADN cible vont être allongées et se lier à la sonde spécifique de locus. Le fragment ainsi formé sert de matrice pour les amorces universelles. Le produit PCR obtenu sera marqué par un chromophore spécifique de chaque allèle

Les sondes utilisées ont été définies avec les caractéristiques suivantes :

- Les sondes ASO possédaient en 5' une séquence correspondant à l'amorce T3 pour l'allèle 1 et à l'amorce T7 pour l'allèle 2. La partie spécifique d'allèle était composée de 22 bases et se terminait en 3' par le SNP. La conception de cette partie était donc très restrictive puisque seules les zones entourant le SNP pouvaient être utilisées. Si celles-ci n'avaient pas un pourcentage en GC correct ou si les séquences

généraient des structures secondaires, le SNP ne pouvait pas être analysé.

- La sonde LSO était composée de trois parties. En 5', nous avons la séquence complémentaire d'une amorce universelle FW. Au milieu, une séquence d'adressage (C-Zip) complémentaire de la séquence de capture présente sur le futur damier a été ajoutée. Cette séquence était unique pour chaque locus analysé. Elle est généralement issue d'une sélection aléatoire dans un génome très éloigné de celui étudié, comme par exemple un bactériophage λ (Hirschhorn et al. 2000). Enfin, en 3', nous avons la partie spécifique du locus positionnée à une dizaine de bases du SNP. Cette partie de la sonde était donc un élément restrictif car les possibilités étaient limitées.

De nombreuses mises au point ont ensuite été réalisées sur le protocole, dont le choix du tampon pour l'extension-ligation, le choix des températures du programme PCR, l'ajout d'une étape de lavage et de précipitation à l'éthanol ou encore le temps d'élongation lors de la PCR universelle.

Un choix important consistait à amplifier ou non au préalable l'ADN cible. Travailler sur de l'ADNg total permettait de diminuer le nombre d'étapes mais augmentait les risques d'hybridations non spécifiques. Nous avons comparé les deux conditions et séquencé les produits d'amplification. Lorsque l'ADN total a servi de matrice (**figure 3.7 A**), beaucoup de bruit de fond a été observé et le fragment attendu (126 pb) n'était pas présent. Après analyse du séquençage des fragments obtenus, ceux-ci ne correspondaient pas à ce qui était attendu. Travailler à partir d'ADNg n'a donc pas permis d'obtenir une spécificité d'amplification suffisante. Lorsque nous avons travaillé à partir de la séquence cible préalablement amplifiée (**figure 3.7 B**), nous avons obtenu les fragments de taille attendue, 121 et 122 pb (avec un léger décalage) sans bruit de fond. Le séquençage des fragments a révélé que sur sept fragments analysés, cinq correspondaient bien à ce qui était attendu. La réaction fut donc concluante dans 70% des cas. L'étape préliminaire d'amplification de l'ADN cible a donc été ajoutée au protocole.

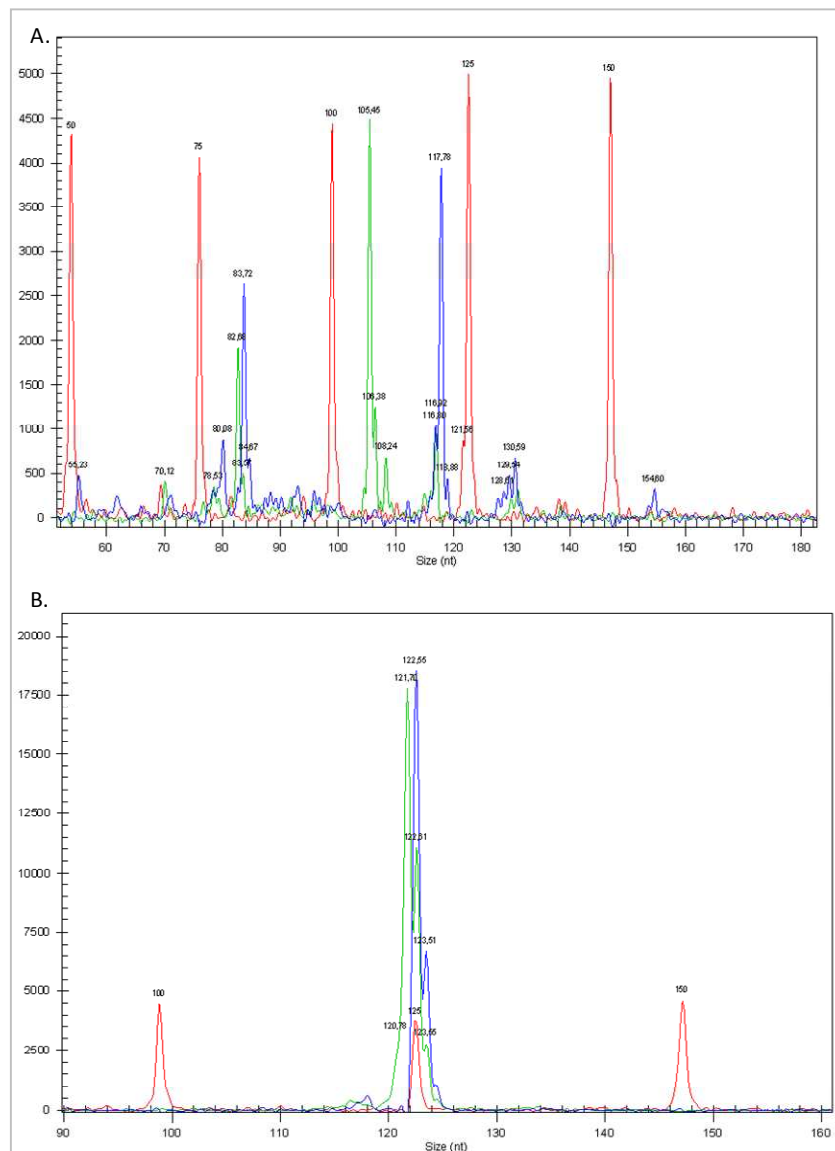


Figure 3.7. Comparaison des résultats de l'extension-ligation réalisée à partir d'ADNg (A) et d'un produit PCR (B). Visualisation de la taille des fragments après migration sur capillaire (bleu et vert : fragments marqués, rouge : standard de taille). Abscisse : taille des fragments en pb. Ordonnée : intensité du signal. Taille du fragment attendu : 126 pb

Les résultats obtenus par la méthode d'extension-ligation se sont donc avérés concluants et spécifiques. Les fragments ainsi obtenus pouvaient être identifiés par migration sur séquenceur capillaire. Bien que prévue lors de la conception des sondes, la détection sur damier n'a pas été mise au point car, au final, le nombre de marqueurs développés pour la cartographie ne nécessitait pas une telle plateforme.

IV.III.1.2. Méthode de ligation

Afin de simplifier le protocole de génotypage et de voir si la ligation seule permettait d'obtenir une spécificité égale par rapport au protocole détaillé ci-dessus, nous avons supprimé l'étape d'extension. Le principe est identique mais la sonde LSO doit se trouver juste à côté du SNP. Seule une hybridation parfaite des sondes ASO et LSO permet une ligation et donc la formation d'un fragment servant de matrice pour l'amplification spécifique (**figure 3.8**).

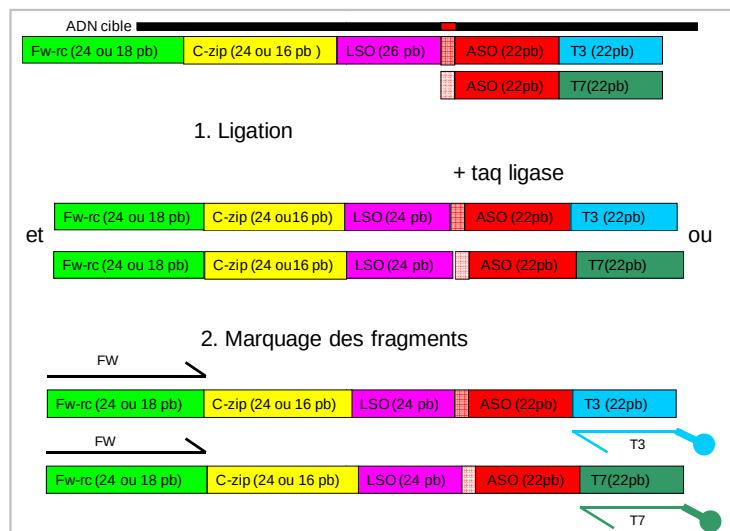


Figure 3.8 : Méthode de ligation. T3, T7 et FW : amorces universelles. ASO : oligo spécifique d'allèle. LSO : oligo spécifique de locus. C-Zip : séquence d'adressage. Fw-rc : « reverse complement » de l'amorce FW.  et  : chromophore.  : emplacement du SNP. Les sondes spécifiques d'allèle, une fois parfaitement hybridées sur l'ADN cible, vont se lier à la sonde spécifique de locus. Le fragment ainsi formé sert de matrice pour les amorces universelles. Le produit PCR obtenu est marqué par un chromophore spécifique de chaque allèle

Plusieurs mises au point ont dû être réalisées, entre autres sur la quantité de ligase, sur les lavages ainsi que sur la quantité de produit de ligation à utiliser pour la PCR universelle. Les fragments obtenus ont été séquencés et correspondaient dans tous les cas, parfaitement à ce qui était attendu. Cette méthode semblait donc plus spécifique que l'extension-ligation.

Ces résultats sont étonnants puisque la méthode d'extension-ligation possède plus d'étapes restrictives (hybridation des sondes, extension et ligation) que la méthode de ligation (hybridation des sondes et ligation). L'extension semble être responsable de la présence des fragments non spécifiques puisque cette étape est la seule qui diffère entre les deux méthodes.

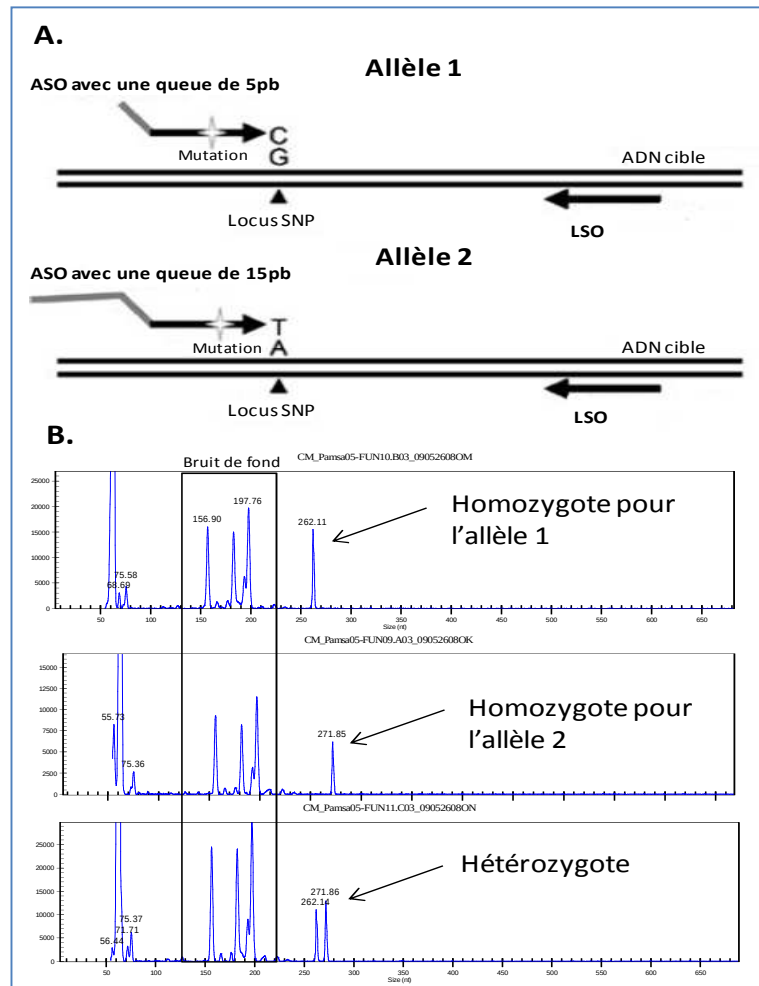


Figure 3.9 : **A.** Principe de la méthode de génotypage PAMSA. L'ADNg est amplifié à l'aide de deux amorces spécifiques d'allèle (ASO) et une amorce spécifique de locus (LSO). Une mutation a été ajoutée sur les amorces ASO afin d'augmenter la spécificité de l'amplification. L'allèle 1 est amplifié à l'aide de l'ASO comprenant une queue de cinq bases et l'allèle 2 est amplifié à l'aide de l'ASO comprenant une queue de 15 bases. Cela génère donc des produits d'amplification de tailles différentes (Gaudet et al. 2007). **B.** Exemple de sorties après migration sur séquenceur capillaire des fragments obtenus par la méthode PAMSA. Abscisse : taille des fragments en pb. Ordonnée : intensité du signal. Les pics compris entre 150 et 200 pb sont dus au bruit de fond généré par l'utilisation de dUTP marqués. Le marqueur PAMSA 5 génère un fragment de 262 pb pour l'allèle 1 et de 272 pb pour l'allèle 2. Les deux homozygotes et l'hétérozygote sont donc facilement reconnaissables

IV.III.1.3. Méthode PAMSA

Afin de pouvoir, dans un premier temps, génotyper à l'aide du séquenceur capillaire et non pas sur damier, une nouvelle méthode, plus adaptée à ce mode d'analyse a été mise au point. Le terme PAMSA signifie « PCR Amplification of Multiple Specific Alleles ». Cette méthode est issue de l'article de Gaudet et al. (Gaudet et al. 2007) (**figure 3.9 A**). La discrimination des allèles s'est faite sur base d'une différence de taille entre les produits PCR obtenus.

Trois amorces ont été définies par SNP : 2 ASO et une LSO. Les deux amorces ASO devaient se terminer en 3' par le SNP. Sur ces amorces, on retrouvait également une étiquette de cinq pb pour l'allèle 1 et de 15 pb pour l'allèle 2. De plus, une mutation propre à chaque allèle a été insérée au niveau de la 4^{ème} base en partant de l'extrémité 3'. Cette mutation permettait d'augmenter la spécificité des amorces ASO. En effet, après le premier cycle d'amplification, deux bases, au lieu d'une seule, étaient discriminantes pour l'hybridation des amorces. La conception de l'amorce LSO n'a pas subi de contrainte particulière.

En présence de l'ADN cible, les amorces ASO s'hybridaient de manière spécifique sur l'allèle correspondant. L'hybridation simultanée de l'amorce LSO permettait à la taq polymérase d'amplifier le fragment correspondant à l'allèle 1 et/ou à l'allèle 2. L'amorce ASO 2 possédant 10 pb de plus que l'amorce ASO 1, il était facile de discerner par migration sur séquenceur capillaire les deux types de fragment (**figure 3.9B**).

Cette méthode présentait plusieurs avantages. Tout d'abord, l'obtention des fragments ne nécessitait qu'une seule PCR. De plus, l'amorce LSO pouvait être positionnée n'importe où sur le locus. Cela laissait donc une marge de manœuvre plus grande par rapport aux méthodes d'extension-ligation et de ligation. Enfin, le marquage des fragments a été réalisé à l'aide de dUTP marqués s'incorporant lors de l'étape d'élongation. Une sonde marquée pour chaque locus n'était donc pas nécessaire, diminuant considérablement le coût des amorces.

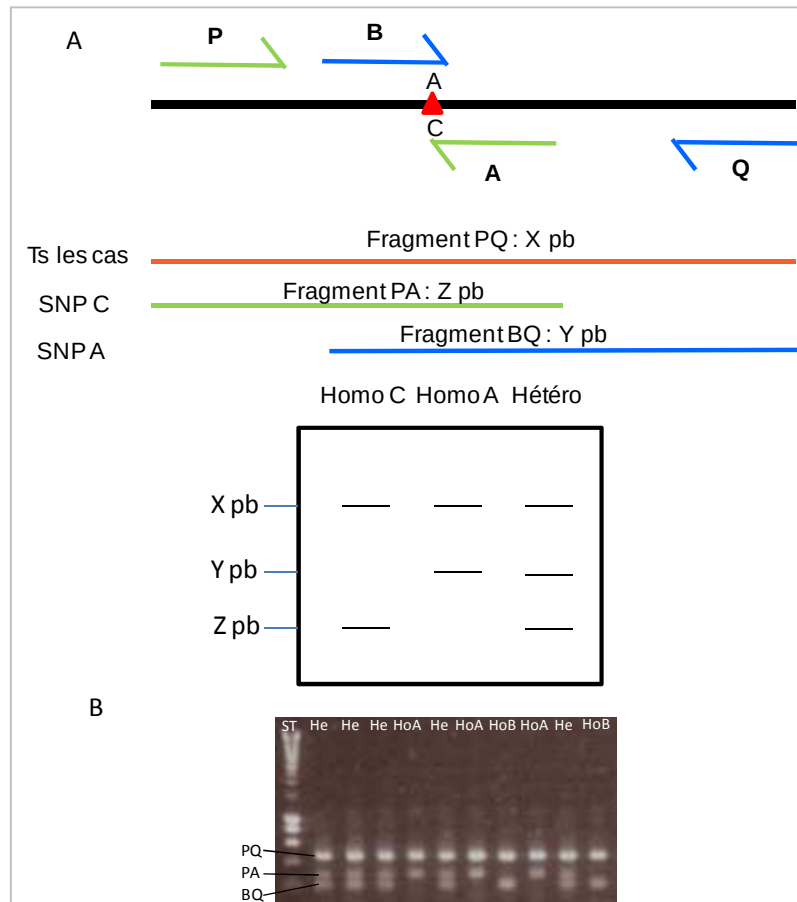


Figure 3.10 : A. Principe de la méthode Bi-PASA. Quatre amorces sont nécessaires par locus analysé. Les amorces P et Q sont spécifiques de locus. Les amorces B et A sont spécifiques d'allèle. L'extrémité 3' de ces deux amorces se termine par le SNP (ici, C pour l'amorce A et A pour l'amorce B) et leur orientation est opposée. En présence du SNP C, le fragment PA de x pb est amplifié. En présence du SNP A, le fragment BQ de z pb est amplifié. L'amplification du fragment PQ est un contrôle positif pour la PCR. La différence de taille entre les différents fragments obtenus permet de discerner les deux allèles sur gel d'agarose et donc d'identifier les individus homozygotes et hétérozygotes. **B.** Exemple de génotypage par la méthode Bi-PASA à l'aide du marqueur bi-pasa test2. Fragment PQ : 509 pb, fragment PA : 312 pb et fragment BQ : 230 pb. ST : standard de taille, He : hétérozygote, HoA : homozygote pour l'allèle 1, HoB : Homozygote pour l'allèle 2

IV.III.1.4. Méthode Bi-pasa

Afin de réduire encore les coûts de génotypage liés à l'analyse des fragments au séquenceur, une méthode détectable sur gel d'agarose a été mise au point. La technique Bi-PASA (Bidirectional PCR amplification of Specific Alleles) décrite par Liu et al. (1997) (Liu et al. 1997) est basée, comme la méthode PAMSA, sur la discrimination des allèles en fonction de la taille des fragments générés.

Quatre sondes ont été déterminées pour chaque locus étudié : deux ASO (amorces A et B) et deux LSO (amorces P et Q) (**figure 3.10**). Chacune des amorces ASO était spécifique d'un allèle au niveau de l'extrémité 3'. La principale différence était qu'une des amorces s'hybridait sur le brin de polarité positive et que l'autre s'hybridait sur le brin de polarité négative. Une étiquette d'une dizaine de bases GC a été ajoutée à l'extrémité 5' de ces deux sondes afin d'augmenter l'efficacité d'amplification des fragments spécifiques d'allèles. L'amorce LSO P permettait l'amplification du fragment PA en présence de l'allèle 1, et l'amorce LSO Q permettait l'amplification du fragment BQ en présence de l'allèle 2. Ces deux fragments ayant des tailles différentes, il était possible de les visualiser sur gel d'agarose. Les amorces PQ généraient un fragment de taille supérieure aux deux autres. Celui-ci servait de contrôle positif pour l'amplification.

Le mode de détection de cette méthode constituait un avantage majeur. En effet, se passer de séquenceur et ne pas utiliser de sonde marquée permettait de diminuer fortement le prix de chaque analyse. Cette méthode a été utilisée pour le génotypage des séquences possédant un orthologue potentiel sur la carte de la laitue.

IV.III.1.5. Récapitulatif des méthodes de génotypage

Le **tableau 3.20** récapitule les caractéristiques de chaque méthode de génotypage de SNP mise au point. Le choix de la méthode de génotypage dépend de plusieurs facteurs. Tout d'abord, le nombre de SNP à analyser va influencer le choix de la méthode de détection des allèles. Ensuite, afin de diminuer la durée des manipulations, le nombre d'étapes doit également être pris en compte. Enfin, l'utilisation ou non d'un marquage à l'aide de chromophores ainsi que la taille des sondes utilisées influence le coût des analyses

Tableau 3.21 : Récapitulatif des sondes PAMSA valides et description de la fonction des séquences associées. Gris : Marqueurs utilisés pour le génotypage des 200 individus F2 du croisement SAMCHIC

Nom du marqueur	Contig	Description de la fonction de la séquence homologue	eValue	Pourcentage de similarité	LSO	ASO 1	ASO2
1	1531	fructose-bisphosphate aldolase	4.31E-53	80.05%	CCAAGGGCATATAACAACACC	GCTGTACAAGAGCGTTTCTGGCA	TAATCATTACCGTGAACAAGAGCGGTTTCTCGCG
2	605	transaldolase	5.03E-86	94.2%	ACGAATCCACTCTTCAACC	GCGGAAGAATCCAGCATACCCGGAC	ACGGTCAGAGAGATTAGAATCCAGCATACCCGAT
3b	1	poly-binding protein	1.25E-41	82.75%	CAAAGAAAGGCAACAATCTCC	GTTATTGCTGTTGGAGATGGATCATACA	GGTCAGTTTGAGCATTGCTGTGGAGATGGATCAACT
4	578	acid invertase	1.17E-86	83.9%	ATGGATCGAGTTGTGCTTCC	TTACCCAAGGTCATCGACGAAAGG	TATGATTGATCGTGCGAAGGTCATCGACGGAATAAT
5	1092	elongation factor 1-alpha	1.81E-80	98.35%	TCACCTTCACTCCCTTCTTGG	GCTGTGACCAAGATTGACAGCGCA	GGCATTGACAGAGATGACCAAGATTGACAGCGCG
6	112	protein	1.45E-77	99.1%	AACCCATCGCATCTTTTGG	TGACGGCTGGTGGAGTGTITCTAGTAACC	TCATTGTAGCGGGAGGCTGGTGGAGTGTITCTAGTTACT
7	1130	---NA---	-	-	TCCCACCATCTATCTACTGC	TCAATCAGTTGTTGAGGAAACCGATAAA	AATCAACGTAAAGCGCAGTTGTTGAGGAAACCGAAAAG
9b	1248	oxygen-evolving enhancer protein chloroplast	1.58E-86	91.0%	AACACAAACGCTCGAGGAAC	CCATTACCAACTGATCACCCTACC	ATCGCCTGGTTCATTACCAACTGATCACCCTAACT
10	1273	40s ribosomal protein s11	1.52E-61	94.7%	CACCACCACCTGAAGATCC	GGCGTCTGCTCACATCTCACCTGCG	AACAGAGCGTAGCCCTGCTCACATCTCACCGCTGT
14	173	s-adenosylmethionine synthetase	1.25E-97	92.9%	GTCTCCAGGTGAAGTCAGG	AACATGCAAGCGGCTAAGAGGATC	GAGTCAATCGCAGACGCAAGCGGCTAAGAGCATT
15	253	14 kda proline-rich protein	3.68E-14	88.55%	GGATGGGGATTCTTTCCAC	CATATGCACAGCCATTAAAGCTCATA	AAGGTACTGCAAGTGGCACAGCCATTAAAGCTGATG
16	1447	cationic peroxidase	1.22E-150	93.25%	GACCGTGGTACCCCTATGAA	CTGCAACCAAGTGGATTGTTATCG	TTCTCGGTAGCACACAGTGAAGTGGATGTTTTCAT
17a	1163	polyubiquitin	5.09E-33	97.75%	ATCTGCATACCTCCACGGAG	AACCTGTTATCGTCTTTCCGGCCAA	AACATTGATTCAAGTGTATTGCTCTTTCCGGGCGAG
18b	1241	ribulose--bisphosphate carboxylase oxygenase	6.17E-80	90.3%	CTGCCCTTCCCATTACAAG	ATATGGTGCTCCAACCTCGAATTCAAAG	TTATCTGCCACGCCGGTGCTCCAACTGAAATCTAAA
20	1423	annexin 2	2.36E-69	85.75%	CAATCCCTCACACAAACCAA	TGGTTCGATTAAACAAACGAACCGAT	TCACGACGTTCTCAGCATTAAACAAACAGGAACGGAC
21	1536	ribosomal protein	6.85E-70	91.7%	TCATCGTCCCTCATCATCAGC	GAAGCGAAGATGAACCTGAGAAGTAGCAA	CCAGACATGGTGGTTGAAGATGAACCTGAGAAGTAGCAG
22	198	glyceraldehyde-3-phosphate dehydrogenase	1.57E-71	94.8%	TGTCCTTCCGTGTACCAACA	ATCTCTGTCATTCAAGGCAATGCCA	CGCTGGACTATCGAAGTCATTCAAGGCAATCCCG
23	235	tubulin beta-2 beta-3 chain	4.44E-68	78.4%	TCAACCGTGTGTGACATTCC	GAACGCCCTCCTCATCAGCAGTAGCG	CAATATTCAACAGCACCTCTCTCATCAGCAGTTGCCA
26	328	at3g27090 moj10_18	6.93E-78	95.55%	CCCCACAAACAACTCTACC	GCAGACTGTTTCGAGCAATTACCCCA	ACTGATGAGCGATCCCTGTTCGAGCAATTACGCCT
27	402	heat shock protein 70	6.89E-92	85.35%	GCGAAGAATGCTTTGGAGAA	GCAGACGCCCTACGCCAATCTTTG	ACTGATGAGCGATCCGCCCTCAGCCCACTGATTT
28	484	endo-beta-glucanase	3.63E-58	89.95%	AGGAATTAAATGGCGGGATG	CCTGTTCCATCCACCACTCATCCAA	GCGAGACAGCGACGATCCATCCACCACTCATGCAC
29	498	---NA---	-	-	CACAAGAAACAGTAGCCGAGTC	AAGGAGTCGCTGCTGCTGCTCTA	GATTATGACCGAGGGTCGCTGCTGCTGACTG
30	500	polyubiquitin	4.68E-108	99.0%	CTCCACCTCCAAGTAATGG	AATACACAACGTCAAGGCCAACATT	CGGAACATCTCGGTAAACAACGTCAAGGCCAAGATC
31	547	proline-rich protein	8.80E-41	60.2%	GGGGTAGTTTTTAAAGGAAAGCAA	CATGGCACCTGTACCAAGTCTGAGTCA	AGGAAGATATCCTCGCACCTGTATACCAGGCTGACTCC
32a	589	ap2 erf domain-containing transcription factor	4.65E-54	56.05%	GAAGCCGAAATCCAGAAAG	TTGCGCCATCATCTTGAGTTGAACCTC	AGGCCAGTCAACAGCCATCATCTTGAGTTGAAGCTT
32b			4.65E-54	56.05%	GAAGCCGAAATCCAGAAAG	TCCAACCCAGTTCCTTCCAAGGTAC	TCATGGATACAGTTCACGTTCCCTCCAAGCTAA
32c			4.65E-54	56.05%	GAAGCCGAAATCCAGAAAG	GGAACTCTTCTCACTGGACTACCTCTCC	GGATATACCGCTCAACCTCTTCACTGGACTACCGTCT
34	704	60s ribosomal protein l7	4.49E-91	94.05%	CAGCATCAACCACTTCAACA	CAGGAAAGAATCCCATTAACAGGCAAT	CCAATCTAGTAACCAAGAATCCATTAAACAGGGAAC
35	754	glyceraldehyde-3-phosphate dehydrogenase	9.62E-43	97.85%	CCAAAACATTCCACCCATC	TGAGCGGTTACACTGAAGATGACGTTGTC	GATGACCATCTACATGGTTACACTGAAGATGACGTTAGT
36a	783	phenylalanine ammonia-lyase	2.12E-81	78.05%	GCAAAAGCTCCGTCAAGTTC	AACGTCTAACAAATCGGAAGAGGGTCC	CAGCGAGCTGGTTGACTAACAAATCGGAAGAGGCTCA
36b			1.00E+00	78.05%	GCAAAAGCTCCGTCAAGTTC	CGCAGCAGCTTCAACACATTCCAAAAC	GAGCTTGTCCATATGCAGCCTTCAACACATTCCAATAAT
37a	942	udp-glucuronate 5-epimerase	2.54E-25	91.05%	TCCTTTGGGCACAAAATCAAT	GAGCCAGAAGCAACATAGAGGGTGC	TGGCTAACCCTGACCAAGAAGCAACATAGAGGGCTGG
37b			2.54E-25	91.05%	CAATCCCTCCATTCTCTCAGA	AACATTACGGGATCAAAACAGAGTG	CGCACAGATGTTAATACGGGATCAAAACAGAGTGA
39	32	5-methyltetrahydropteroyltriglutamate-homocysteine	1.99E-68	93.7%	CTCCAAAACGCAAGCATT	TTGCGAAGTACGGTGCTGTTATCGGG	TTGAGGCTTATTTCGGAAGTACGGTGCTGTTATGGGA
41	467	pathogenesis-related protein 10	5.48E-72	66.75%	AGCTTCAGGATGAGCAATGG	AATCCACCATCTTTGAAGGTGATGATTT	CTTTGAACTGAGAAGCACCATCTTTGAAGGTGATGTTTTA
42	934	catalase	0	90.9%	CGTGGCCACATAGTTTATAGG	TGCTTCGTTTGCACAAATGGGGTC	GAATCGAGTGCGCTGCTTTGCAACAAATGGCGTT
43	968	elongation factor	1.01E-49	95.25%	CCGCCAATCCTTTTCTCTCT	ACTATAAGGCGTATCTCTGTCGTT	ATCTAACACCGTCCGAAGGCGTATCTCTGTGGTC
44	1065	class I chitinase	3.23E-61	84.9%	ACATGTGCGGAGCAAAAAAC	CACTTAAACCGCACCATTCAAACT	CTAAGTGACGGCTGCAACCGCACCATTCAATCAG
46	1399	s-adenosyl-l-homocysteine hydrolase	1.20E-49	96.7%	CCAAACACCCCAACAAACTT	GCTGTAGCTATGGAAGGAAGAGGCC	AATGATTCTGTCATCTAGCTATGGAAGGAAGAGCCCA

Tableau 3.20 : Récapitulatif des caractéristiques des quatre méthodes de génotypage de SNP

	Extansion-Ligation	Ligation	PAMSA	Bi-PASA
Méthode de discrimination des allèles	Extension d'amorce et hybridation spécifique	Hybridation spécifique	Extension d'amorces	Extension d'amorces
Mode de détection	Damier et séquenceur	Damier et séquenceur	Séquenceur	Gel d'agarose
Marquage fluorescent	Sondes universelles	Sondes universelles	dUTP marqués	Aucun
Nombre total d'étapes	5	5	2	2
Contraintes liées à la conception des sondes	++	+++	+	+
Taille des sondes (base)	64	64	35	30

IV.III.2. Analyse des SNP présents au sein des séquences de chicorée

Dans ce point, nous allons détailler le génotypage de la population SAMCHIC à l'aide des méthodes de génotypage détaillées ci-dessus pour les SNP mis en évidence lors du chapitre 2.

Les SNP mis en évidence dans les EST de la banque Phytomol ont été analysés à l'aide de la méthode PAMSA. La méthode Bi-PASA a été utilisée pour les SNP restants (EST orthologues de laitue et séquences supplémentaires). Enfin, l'Ecotilling a été utilisé pour l'analyse de séquences impliquées dans la synthèse de l'inuline.

IV.III.2.1. Analyse des SNP présents au sein des séquences EST de la banque Phytomol

Dans le chapitre 2 point II.III, 294 SNP contenus dans 64 contigs ont été mis en évidence au sein des séquences EST de chicorée. Des amorces PAMSA ont été développées afin d'analyser ces SNP au sein du croisement. 53 sondes spécifiques de SNP présents au sein de 47 contigs ont pu être définies. Une partie des SNP n'a pas pu être analysée car les contraintes liées à la conception des amorces PAMSA ne le permettaient pas. De plus, une seule sonde a été définie par haplotype. En effet, l'analyse d'un seul SNP a suffi pour connaître l'identité des autres SNP de cet haplotype. Ces deux raisons expliquent donc pourquoi seuls 18% des SNP ont été utilisés pour le génotypage.

Après mises au point, 39 sondes PAMSA s'avèrent valides. Le **tableau 3.21** reprend les caractéristiques de chaque sonde ainsi que la fonction associée via l'analyse Blast2Go à la séquence contenant le SNP.

Tableau 3.22 : Récapitulatif des marqueurs Bi-PASA fonctionnels. Gris : Marqueurs utilisés pour le génotypage des 200 individus F2 du croisement

Nom du marqueur	Origine	Nom de la séquence	Numéro d'accèsion	Description de la fonction de la séquence homologue	Amorce A	Amorce B	Amorce P	Amorce Q
bi-pasa 02	orthologue laitue	CCIL10409.B1_A11	EH672777	TIR-NBS-LRR resistance-like protein	GGGGCGGGGCATCTTGCTTCCTGGATA	GGGGCGGGGCTGGAACAAAGGTTTTC	TGCAAAACAGCTTCTAGGG	GTTTGAAACCCGCATTCTC
bi-pasa 03	orthologue laitue	CCIM9953.B1_B18	EH698135	Beta-fructofuranosidase	GGGGCGGGGCACTCCACAATAACAATG	GGGGCGGGCGCAGAACATTTGTAATCCG	CTCTTGAAGGTCAAGAATGGC	CTATCTTCACTGCTCTTTCTGCA
bi-pasa 20	orthologue laitue	CCIM3748.B1_H02	EH691568	Grave disease carrier protein	GGGGCGGGGCGCTTGTTCTGTTACGC	GGGGCGGGGCGGTAGTTTGTCAATTACA	ACCATTGCGGTTAGTTGAGG	ATCCTTATTGGGCGATGTTG
bi-pasa 21	orthologue laitue	CCIS4138.B1_D04	EH705277	Rubisco subunit binding-protein alpha subunit	GGGGCGGGGCTCCAACATATCACCCTAA	GGGGCGGGGCATGCTATATGATTTGTGAGTG	TGCCATTGAAGAAGGATTG	TAAATCGCATGTGTGCTGTG
bi-pasa 23	orthologue laitue	XCI2_CONTIG2197	EH682079	Adenylosuccinate synthetase	GGGGCGGGGCGGATCATGTAATGCGATG	GGGGCGGGGCGGCAATTGAATTTAGCC	CACAACACCATCAGTGGAG	GATCAAGCTTTTGAGGAAGG
bi-pasa 28	orthologue laitue	XCI2_CONTIG6915	EH709102	ran-binding protein 3-like (LOC100082419)	GGGGCGGGGCTTGGGGAMCGTATGGA	GGGGCGGGGCGAGTCTTTCGCCGCAC	CCTTTCGATAAAACGGAAGATG	TCTCCTWTACGCCTCCAGAC
bi-pasa 33	orthologue laitue	CCIL10281.B1_B04	EH672649	Elicitor inducible beta-1,3-glucanase NIEIG-E76	GGGGCGGGCGCGATTGCGTCGATGAA	GGGGCGGGGCTACCACCTTCCACCA	TGCAGATAACCTCCCTGACC	CCGGTTTTGGATCTTTCAC
bi-pasa 37	orthologue laitue	XCI2_CONTIG2587	EH684149	ethylene-insensitive 2	GGGGCGGGGCAATGGAATCATTAAACCG	GGGGCGGGGCATGACGGTGATTTGGG	ATCTTTCGGTTGCATTGAG	GTCAAAGTTAGCGGGATTGG
bi-pasa 41	orthologue laitue	XCI2_CONTIG792	EL345653	coniferylalcohol 5-hydroxylase	GGGGCGGGCGCAGCGGAACCGGAG	GGGGCGGGCAATCACCCTGCAAGT	AAAATGGGTTTCAGCCACAC	AAACGCCAACATCTCATCC
bi-pasa 43	orthologue laitue	CCIS6886.B1_L18	EH707518	predicted protein	GGGGCGGGGCGGGCAAGCGAATATC	GGGGCGGGGCGAGCGTTTTTGCT	TGAACACCCACCACACATA	CGTCGTGCTGTTCTCATAG
bi-pasa 44	orthologue laitue	CCIM1833.B1_B04	EH689663	predicted protein	GGGGCGGGGCAAGGAATATAGTTTCTCC	GGGGCGGGGCTCCAGGTACAGCACCA	AACAAGGTGCACCAAAACC	CAACCGTGTAGCCATAGCAG
bi-pasa test2	orthologue laitue	CCIL1586.B1_C13	EH676972	barley mlo defense gene homolog6	GGGGCGGGGCGTCAAAACAGGCAAGA	GGGGCGGGGCTCGTGGTAATTAACAAG	CGTAAGTATCTTCGATACACAAAC	GGAGGTAGAGGTCTAGATGAAC
bi-pasa test4	orthologue laitue	CCIS14137.B1_B08	EH702104	-	GGGGCGGGGCGCCAGCAGATCTCGTGATC	GGGGCGGGGCTCGAAAAGACGTCTACA	CCTCAAAAGAAGACGTACACA	TCATCTCCACCACTTCATCAT
bi-pasa34	orthologue laitue	XCI2_CONTIG5358	EH697036	-	GGGGCGGGGCGGTGAAAAGCCGA	GGGGCGGGGCGCCACAACATAATCAACG	TCCGACACATGAGGATCTTG	ATATACTGCCGATGGTGACG
bi-pasa35	orthologue laitue	XCI2_CONTIG792	DW096634	ferulate 5-hydroxylase (F5H-K10)	GGGGCGGGGCTGGGATAAAATCGGCTAA	GGGGCGGGGCTTTTGGTGCTTCAACC	AAAATGGGTTTCAGCCACAC	AAACGCCAACATCTCATCC
bi-pasa-49	STS cadalen	MADs box L2/R2	AF101420	MADS box protein	GGGGCGGGGCGCGGTATCCTTATGTAGCTC	GGGGCGGGGCGCTCTTGGCTTCCACA	AAGCCTTGAAGGTCTCTCC	AAAGTTGACGAACCTTCTCTCG
bi-pasa-53	STS cadalen	AF497999 L2/R2	AF497999	Germacrene A synthase long form	GGGGCGGGGCTAGGGTTCTTTCAGACG	GGGGCGGGGCGCAGGACTTAGATCCG	TATCCAGCAGTCGTTCCAGA	GTCACACCGTCGAGAAAGG
bi-pasa-57	floraison	ciflw12	EH692696	transcription factor IIIA (TFIIIA)	GGGGCGGGGCAAGCTTCTTTCATCTCC	GGGGCGGGGCGAGTGTTTTGGGAATACA	GGATTAGATGGCGTTTAGG	AAACTCGCACATTAAAGCAG
bi-pasa-59	floraison	Gladng	EH682796	G10 GIGANTEA gene	GGGGCGGGGCGAGCTCACACGCGAGA	GGGGCGGGGCGCGGAGTTTGTGCA	GGGGGTAAGGTTAAGAGATG	ACTCCATGAAGTCCCGACAG
bi-pasa-66	pathogénese	CIC.EH691923	EH691923	wall-associated kinase (WAK)	GGGGCGGGGCAACATCGCTCTTGTGCG	GGGGCGGGGCGCAATGCTACCAGTTGACT	AAACAGCCGAAGCACTAAC	GAATCCCCCTCAAATCTCC
bi-pasa-70	pathogénese	Contig1501	FL679098	CBL-interacting protein kinase	GGGGCGGGGCGGAGCGGTTGTC	GGGGCGGGGCGCTTCTCCAGCTTCGAC	GCCAGACGATGCTCTAAGC	CCAAGTCTTGTGGAAGACG

Afin de vérifier si les marqueurs SNP étaient polymorphes dans le croisement, huit individus de la F2 ont été génotypés à l'aide des 39 sondes. Au final, 6 marqueurs ayant une distribution 2:1:1 (test χ^2 , $p>0.01$) ont été retenus et utilisés pour le génotypage des 200 individus de la F2 (lignes grises du **tableau 3.21**). Cela représentait 15% des marqueurs analysés. Cette valeur plus faible que celle obtenue pour les microsatellites (24%) peut s'expliquer par le fait que certains marqueurs ont été éliminés car les produits d'amplification semblaient non spécifiques.

IV.III.2.2. Analyse des SNP présents au sein de séquences orthologues à la laitue et de séquences d'intérêt (carte Cadalen, pathogénèse, floraison, ...)

Du polymorphisme SNP a été mis en évidence au sein de 94 séquences choisies (Chapitre 2 point III et IV et **tableau 2.7**). Des sondes Bi-pasa ont été définies pour 74 d'entre elles. Chaque combinaison de sondes (PA, QB, POA, PQAB) a été testée sur quatre ADN de chicorée et un mélange de huit ADN provenant d'individus du croisement afin de vérifier la fonctionnalité des amorces ainsi que le polymorphisme de ces marqueurs dans la F2. 51 marqueurs ont ainsi été retenus.

Ces marqueurs ont été utilisés pour amplifier 16 individus du croisement. Après analyse des profils, 30 ont dû être écartés car ils présentaient des génotypes non spécifiques. En effet, soit les individus étaient tous hétérozygotes, soit un des allèles était présent chez tous les individus. Cela peut s'expliquer par le fait que certaines amorces avec une température d'hybridation trop faible, génèrent une amplification, quel que soit l'allèle présent. Des mises au point ont été réalisées pour en augmenter la spécificité, mais certains marqueurs ont tout de même dû être écartés.

Au final, 21 marqueurs ont été utilisés pour génotyper les 200 individus du croisement (**tableau 3.22**). Cinq marqueurs ne répondant pas à la distribution attendue 2:1:1 (test χ^2 , $p<0.01$) ont été supprimés de l'analyse. Au total, 17 marqueurs pouvaient être cartographiés.

Parmi ceux-ci, 12 marqueurs étaient associés à des orthologues potentiels de séquences positionnées sur la carte génétique de la

laitue. Ces marqueurs ont donc servi de point d'ancrage pour une cartographie comparée entre la chicorée et la laitue (chapitre 5). Deux marqueurs étaient associés à la floraison et deux autres à la pathogénèse.

Concernant les séquences provenant de la carte génétique publiée par Cadalen (Cadalen et al. 2010), un seul STS s'est avéré utilisable dans notre croisement. Cela peut s'expliquer principalement par le fait que la moitié des séquences ne présentait pas de polymorphisme au sein de notre croisement. De plus, les amorces publiées dans l'article généraient des fragments de petite taille. De ce fait, il était difficile de définir des sondes autour d'une grande partie des SNP. Vu les résultats obtenus, il n'a donc pas été possible de faire le lien entre la carte décrite dans ce travail et la carte consensus de Cadalen.

IV.III.2.3. Analyse du polymorphisme au sein de séquences impliquées dans la synthèse de l'inuline

L'inuline étant le principal débouché de la culture de la chicorée industrielle, il semblait important de positionner sur la carte des enzymes directement ou indirectement impliquées dans le métabolisme de ce polysaccharide (**figure 1.3**, chapitre 1). Huit séquences ont ainsi été analysées au sein de la population F2 afin d'évaluer leur polymorphisme :

- 1-SST : sucrose-1-sucrose fructosyltransferase, l'enzyme responsable de l'initiation de la synthèse de l'inuline
- 1-FFT : fructose-1-fructose fructosyltransferase, l'enzyme responsable de l'augmentation du degré de polymérisation de l'inuline
- 1-FEH I, II-a et II-b: trois fructose exohydrolases, des enzymes responsables de la dépolymérisation de l'inuline
- Cifr03, cifr07 et putinv : trois invertases impliquées dans la dégradation du saccharose en glucose et fructose.

Les amorces et les conditions d'amplifications spécifiques de chaque séquence ont été fournies par N. Dauchot (Dauchot 2007).

Deux autres séquences (est18 et est40) ne présentant pas de polymorphisme SSR dans le croisement ont également été analysées.

Suite à l'analyse Ecotilling de ces 10 séquences chez l'individu F1 du croisement SAMCHIC, il s'est avéré qu'est18, est40, FEH1 et Cifr07 présentaient du polymorphisme. Ces quatre séquences ont dès lors été analysées par Ecotilling chez les 200 individus de la F2. Ces 4 marqueurs présentaient une distribution attendue (3 : 1, test χ^2 $p < 0.01$) et pouvaient donc être positionnés sur notre carte génétique.

IV.III.2.4. Récapitulatif des SNP analysés pour le génotypage

Le **tableau 3.23** reprend l'ensemble des marqueurs SNP développés dans cette analyse.

Tableau 3.23 : Récapitulatif du nombre de séquences analysées et des méthodes utilisées pour le développement de marqueurs SNP. Le nombre de marqueurs valides ainsi que le nombre de marqueurs utilisés pour le génotypage de la population SAMCHIC sont également indiqués

Provenance des SNP	Nombre de séquences analysées	Méthode de génotypage utilisée	Nombre de marqueurs valides	Nombre de marqueurs polymorphes dans le croisement
Banque EST Phytomol	64	PAMSA	39	6
EST orthologues de la laitue	56	Bi-PASA	15	12
STS de la carte de Cadalen, gènes impliqués dans la floraison et la pathogénèse	38	Bi-PASA	6	5
Gènes candidats pour la synthèse de l'inuline	10	Ecotilling	-	4
Total	168		60	27

Suivant l'origine des SNP, on a pu observer un pourcentage variable du nombre de marqueurs polymorphes dans le croisement. En effet, 15% des marqueurs PAMSA ont pu être retenus pour la cartographie contre 81% pour les marqueurs Bi-PASA. Cette différence s'explique simplement par la provenance des SNP. En effet, les SNP analysés en PAMSA ont été mis en évidence au sein de la banque EST Phytomol, sans aucun lien avec le croisement, alors que les SNP analysés en Bi-PASA ont été présélectionnés via l'alignement des allèles provenant des individus du croisement SAMCHIC. Il est donc logique qu'une grande majorité de ces marqueurs soit polymorphe au sein de la F2.

V. CONCLUSION

Lors de ce chapitre, des marqueurs de type AFLP, SSR et SNP ont été mis au point afin de les utiliser pour la construction d'une carte génétique (**tableau 3.24**)

Tableau 3.24 : Récapitulatifs des marqueurs AFLP, SSR et SNP valides et polymorphes dans le croisement SAMCHIC

Type de marqueurs	Nombre de sites mis en évidence	Nombre de séquences possédant ces sites	Marqueurs valides	Marqueurs polymorphes dans le croisement
AFLP	-	-	-	13
SSR	109	109	70	14
SNP	584	168	60	27
Total	693	277	130	54

L'AFLP a été choisi dans le but de générer un grand nombre de marqueurs sans connaissance préalable des sites polymorphes. Les banques EST de chicorée n'étant à l'époque pas disponibles, l'AFLP paraissait un moyen efficace pour l'analyse du croisement SAMCHIC en vue de saturer la carte génétique. Certains problèmes de génotypage ont cependant été rencontrés, probablement à cause de la méthode de visualisation des fragments d'amplification (séquenceur capillaire). Un grand nombre de marqueurs ont ainsi dû être écartés. Au final, 13 marqueurs AFLP ont été retenus pour être positionnés sur la carte. L'analyse du même croisement avec d'autres paires d'amorces par Claire-Noel Thienpont de l'ECAV à l'UCL a permis d'obtenir 163 marqueurs AFLP (visualisation des fragments via un séquenceur Li-Cor). Au total, 176 marqueurs AFLP ont ainsi été développés pour la cartographie.

Septante marqueurs microsatellites issus de séquences EST ont également pu être développés. Ces marqueurs sont valides et peuvent être utilisés pour l'analyse des variétés de chicorée. Parmi ceux-ci, 14 SSR ont montré du polymorphisme au sein du croisement SAMCHIC avec une distribution 2:1:1 et ont donc été eux aussi retenus pour être positionnés sur la carte génétique. 27 SSR non-géniques ont également été analysés pour le génotypage des individus du croisement (analyses réalisées par la société Chicoline) et ont été utilisés pour la cartographie.

Concernant les SNP, 60 marqueurs PAMSA et Bi-PASA ont été mis au point et sont valides. Au final, 27 SNP ont pu être positionnés sur la carte. Parmi ceux-ci, 12 marqueurs étaient positionnés au sein d'orthologues potentiels d'EST cartographiées chez la laitue. Une cartographie comparée entre les deux espèces a donc été envisagée.

Chapitre 4

Construction d'une carte génétique de la chicorée

Chapitre 4. CONSTRUCTION D'UNE CARTE GÉNÉTIQUE DE LA CHICORÉE

I. BUT DU CHAPITRE

Le but de cette partie du travail était de construire une carte génétique à l'aide des 244 marqueurs AFLP, SSR et SNP génotypés sur la descendance F2 du croisement SAMCHIC (**figure 4.1**). La présence sur cette carte de marqueurs associés à des séquences exprimées a permis de réaliser une cartographie comparée avec la laitue (chapitre 5).

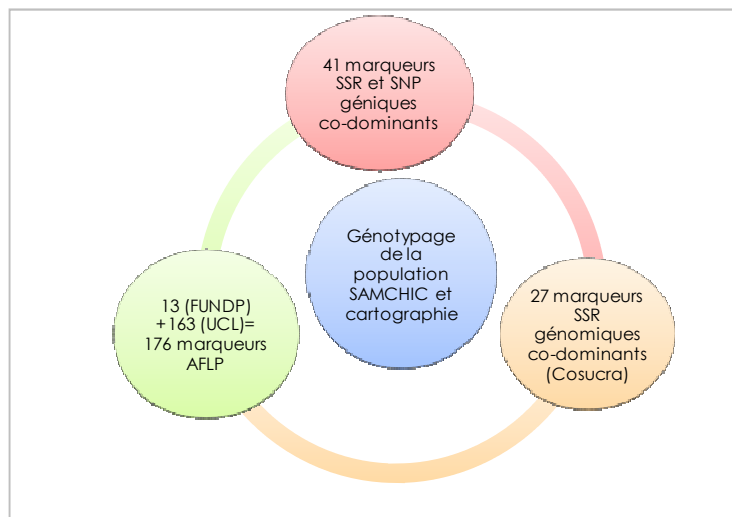


Figure 4.1 : Récapitulatif des marqueurs utilisés pour le génotypage de la population SAMCHIC et pour la construction de la carte génétique.

Comme expliqué dans le chapitre 3, la construction de cette carte génétique est issue d'une collaboration entre trois partenaires. La société Cosucra a créé la population de ségrégation SAMCHIC et a également génotypé 27 marqueurs SSR non-géniques sur ce croisement. Claire-Noël Thienpont (ECAV-UCL) a développé plus de 90% des marqueurs AFLP de cette carte. Le développement de marqueurs SNP et SSR associés à des séquences exprimées a été réalisé dans le cadre de cette thèse (chapitres 2 et 3). Enfin, la construction de la carte en tant que telle a été effectuée de façon conjointe entre l'UCL et notre laboratoire.

Dans ce chapitre, nous allons tout d'abord évoquer les différentes cartes déjà publiées pour le genre *Cichorium*. Ensuite, la construction de la carte ainsi que ses caractéristiques seront détaillées.

II. CARTES GÉNÉTIQUES DU GENRE CICHORIUM PUBLIÉES.

Trois cartes génétiques ont été publiées pour le genre *Cichorium* (Cadalen et al. 2010; De Simone et al. 1997; Van Stallen et al. 2003). Le **tableau 4.1** reprend les principaux critères de chacune des cartes.

▪ Carte génétique de De Simone (1997)

Cette carte a été construite à partir d'une population de 46 individus F1 issus du croisement de deux parents hétérozygotes provenant de l'espèce « radicchio » (*C. intybus* var. *sylvestre*) et de l'espèce *C. endivia* var. *latifolium* (l'endive). Les marqueurs utilisés sont tous anonymes et dominants (RAPD, AFLP et SAMPL). La construction de la carte a été réalisée à l'aide du logiciel MapMaker. Les marqueurs ont été rassemblés dans des groupes de liaison avec un LOD de 4. Au final, 371 marqueurs sont répartis sur 13 groupes de liaison ayant une taille totale de 1.330 cM.

Dans le croisement utilisé par De Simone, le parent faisant partie de l'espèce *C. intybus* est hétérozygote pour 48% des marqueurs. Ce taux très élevé peut s'expliquer par le fait que l'espèce radicchio est issue d'un croisement spontané entre la chicorée rouge et l'endive. L'origine

interspécifique de cette espèce pourrait donc expliquer la grande variabilité observée au sein des marqueurs.

Pour certains LG, le taux de recombinaison entre les marqueurs semble relativement faible. Les auteurs émettent l'hypothèse que l'utilisation d'un croisement interspécifique pourrait être responsable de ce phénomène. En effet, la présence de réarrangements chromosomiques entre les deux espèces empêche la formation de crossing-over dans certaines régions, ce qui diminue le taux de recombinaison.

Tableau 4.1 : Récapitulatif des cartes génétiques publiées pour le genre *Cichorium*

		De Simone et al. 1997	Van Stallen et al. 2003	Cadalen et al. 2010
Croisement		interspécifique	intraspécifique	carte consensus
Espèce		<i>C. intybus</i> var. <i>syvestre</i> x <i>C. endiva</i>	<i>C. intybus</i> L. var. <i>foliosum</i>	<i>C. intybus</i> var. <i>foliosum</i> et <i>sativum</i>
Population de cartographie		46 individus F1	565 individus F2	consensus map
Marqueurs cartographiés	RAPD	16	126	0
	AFLP	283	0	0
	SAMPL	72	0	0
	SSR	0	0	431
	SNP	0	0	0
	STS	0	0	41
	Total	371	126	472
Marqueurs transférables		0	0	472
Groupe de liaison		13	9	9
Taille de la carte		1330	609,6	878
Estimation de la taille du génome		1405	730	912
% du génome couvert par la carte		94,7	83,5	96,2
Distance moyenne entre les marqueurs		3,6	4,7	1,8
Taille moyenne des groupes		86,9	67,7	97,5

■ Carte génétique de Van Stallen (2003)

Le croisement utilisé pour la cartographie a été réalisé à l'aide de deux parents issus de deux lignées fixées de *C. intybus* var. *foliosum* possédant des caractéristiques distinctes. Les 565 individus de la F2 proviennent d'une autofécondation d'un seul individu F1. Pour la construction de la carte, un LOD de 4 a été utilisé afin de rassembler les 126 marqueurs RAPD en groupes de liaison. Au final, neuf groupes de liaison ont été formés générant ainsi une carte de 609,6 cM. La petite taille de cette carte peut s'expliquer par le fait que les parents utilisés pour le croisement sont peu polymorphes. Cela génère peu de recombinaison entre les séquences homologues, ce qui implique des distances plus faibles entre les marqueurs. Cette carte est composée uniquement de marqueurs non transférables

- Carte génétique de Cadalen (2010)

Cadalen et al. ont publié en 2010 une carte consensus à l'aide de marqueurs communs entre trois cartes génétiques de chicorée. La première carte a été réalisée à partir de 177 individus F1 provenant d'un croisement (K28K59) de deux individus hétérozygotes appartenant à la variété hongroise « kospool » de chicorée industrielle. Le croisement utilisé pour la deuxième carte a été réalisé entre un male stérile (MS8) de chicorée industrielle et une population de la variété « cassel » issue de la même espèce. Un seul individu de la population F1 a été autofécondé, générant ainsi une population F2 de 96 individus (population Rubis118). La troisième population de cartographie est issue d'un croisement entre deux parents issus de deux lignées fixées de chicorée witloof (*C. intybus* var. *foliusum*). Un seul individu a été autofécondé générant ainsi une population F2 de 145 individus (population BR).

Les trois cartes ont été construites à l'aide de SSR obtenus grâce à plusieurs banques d'ADNg enrichies. Des STS issus d'EST, d'ADNc et de gènes ont également été cartographiés. Le polymorphisme de ces séquences a été analysé grâce à la méthode SSCP (Single Strand Conformation Polymorphism). Seuls les marqueurs présents dans au moins deux cartes ont été utilisés pour la construction de la carte consensus. Au final, neuf groupes de liaison ont été formés à l'aide de 472 marqueurs (431 SSR et 41 STS) avec une taille totale de 878 cM.

Tous les marqueurs présents sur cette carte consensus sont potentiellement transférables sur d'autres populations de chicorée. Afin de faire le lien entre notre carte et celle de Cadalen, nous avons, dans les chapitres 2 et 3, testé le polymorphisme des STS au sein de notre population F2. Suite à la suppression des séquences non polymorphes, des séquences trop courtes pour la conception des sondes et des marqueurs ne répondant pas à une distribution 2:1:1 dans notre population, un seul marqueur a pu être cartographié. Il n'a donc pas été possible de faire concorder cette carte avec la nôtre.

Au final, les marqueurs positionnés sur les cartes génétiques du genre *Cichorium* ne peuvent être utilisés dans notre croisement car, soit ils sont anonymes, soit ils ne sont pas polymorphes ou valides dans notre population. Nos propres marqueurs ont donc dû être positionnés sur une nouvelle carte afin de disposer d'un outil fonctionnel pour les variétés étudiées.

Le point suivant détaille les protocoles utilisés pour l'encodage des données, la cartographie ainsi que l'estimation de la taille totale du génome en cM. Nous discuterons des différents résultats obtenus dans le point 3.

III. MATÉRIEL ET MÉTHODES

III.I. JEU DE DONNÉES

Tous les marqueurs utilisés pour la cartographie possédaient une distribution mendélienne 3:1 pour les marqueurs dominants et une distribution 2:1:1 pour les marqueurs co-dominants (test χ^2 , $p < 0.01$).

Le génotype de chaque marqueur a été encodé à deux reprises : une fois comme si l'allèle 1 provenait du parent A et l'allèle 2 du parent B et une deuxième fois comme si l'allèle 1 provenait du parent B et l'allèle 2 du parent A. Cette méthode développée par l'ECAV-UCL sera discutée plus bas.

III.II. CONSTRUCTION DE LA CARTE GÉNÉTIQUE

Le logiciel MapMaker 3.0 (Lincoln et al. 1993) a été utilisé pour la construction de la carte génétique. L'unité par défaut est le cM Haldane. Voici les principales étapes :

- Les marqueurs ont été rassemblés en groupes de liaison (LG) à l'aide de la fonction « suggest subset » avec un LOD de 7 et une distance maximale de 40 cM.
- Les marqueurs hautement informatifs de chaque LG (génotype pour minimum 225 des 247 individus de la population SAMCHIC et distance minimum de 5 cM) ont été ordonnés pour créer une carte « framework » via la fonction « order ». La position des marqueurs a ensuite été vérifiée à l'aide de la commande « ripple ».
- Les marqueurs non informatifs de chaque LG ont ensuite été ajoutés au framework à l'aide de la fonction « build ».
- Les marqueurs non associés à un groupe de liaison lors de la première étape ont été assignés à un LG avec un LOD de 4 et une distance maximale de 45 cM. Ces marqueurs ne sont pas intervenus dans le calcul des distances.

La conception graphique de la carte a été réalisée à l'aide du logiciel MapChart (Voorrips 2002).

III.III. ESTIMATION DE LA TAILLE DU GÉNOME

Deux méthodes de calcul ont été utilisées pour estimer la taille du génome en cM à partir de la taille de notre carte génétique :

- Méthode de Fishman (Fishman et al. 2001) :
 $G_e = G_o + (2tG_o)/n$ où G_o est la taille observée de la carte génétique (en cM), t représente le nombre de LG et n , le nombre d'intervalles entre les marqueurs.
- Méthode 4 de Chakravarti (Chakravarti et al. 1991) :
 $G_e = \sum G_{oi} [(k_i + 1) / ((k_i - 1))]$ où G_{oi} représente la taille observée du LG_i et k_i , le nombre de marqueurs sur ce groupe

IV. RÉSULTATS ET DISCUSSION

IV.I. ENCODAGE DES DONNÉES GÉNOTYPIQUES EN VUE DE LA CARTOGRAPHIE

Lors de cette étude, c'est le logiciel MapMaker (Lander et al. 1987) qui a été choisi. Ce programme utilise un algorithme qui permet une analyse multipoints simultanée d'un nombre quelconque de loci. Ce programme est très largement utilisé pour la cartographie faite à partir de grandes populations F2. De plus, de nombreuses fonctions telles que « ripple » (modifications locales), « drop marker » (modification au niveau du groupe) ou encore « error detection » (recherche de doubles crossing-over proches) sont disponibles afin de vérifier la qualité de la carte génétique.

Lors de l'encodage du génotype de chaque individu pour chaque marqueur, il a fallu indiquer si l'allèle provenait du parent A ou du parent B. Dans notre cas, nous ne disposons plus de l'ADN des parents utilisés pour le croisement. Nous ne connaissons donc pas l'origine de chaque allèle. Afin de contourner ce problème, chaque marqueur a été encodé deux fois : une première fois où l'allèle 1 était censé provenir du parent A et l'allèle 2 du parent B et une deuxième fois où l'allèle 1 était censé provenir du parent B et l'allèle 2 du parent

A. Pour le premier encodage, le nom du marqueur restait inchangé. Pour le deuxième encodage, le nom du marqueur était précédé d'un « x ». Au total, nous avons donc le double de marqueurs dans notre fichier d'entrée (**figure 4.2**).

Encodage des données			
	individu 1	individu 2	individu 3
Marqueur 1	A	B	H
xMarqueur 1	B	A	H
Marqueur 2	H	B	A
xMarqueur 2	H	A	B
Marqueur 3	A	H	B
xMarqueur 3	B	H	A

Formation des LG "jumeaux"	
LG1	LG2
Marqueur 1	xMarqueur 1
xMarqueur 2	Marqueur 2
xMarqueur 3	Marqueur 3

Figure 4.2 : Exemple d'encodage des données génotypiques de la population F2 pour l'analyse MapMaker. A : homozygote pour l'allèle A provenant du parent 1. B : Homozygote pour l'allèle B provenant du parent 2. H : Hétérozygote. La provenance des allèles n'étant pas connue (parents du croisement non disponibles), chaque marqueur a été encodé à deux reprises afin de représenter les deux cas de figure possibles : allèle 1 provenant du parent 1, allèle 2 provenant du parent 2 et inversement. La dualité des données a généré des groupes de liaison jumeaux dont un seul a été sélectionné aléatoirement pour la carte finale

Lors de l'assignation des marqueurs dans un groupe de liaison, nous avons obtenu des LG jumeaux possédant les mêmes marqueurs mais avec le nom sans « x » d'un côté et le marqueur annoté avec le « x » dans l'autre LG. Vu la redondance des données, un seul des deux LG a été retenu. Cette technique a donc permis d'encoder les données sans connaître la provenance des allèles.

Le fichier d'entrée du logiciel MapMaker comportait au final le génotype de 247 individus pour 244 marqueurs doublés (176 marqueurs AFLP, 27 EST-SNP, 12 EST-SSR et 29 SSR non géniques) (**tableau 4.2**).

Tableau 4.2 : Nom et type de marqueurs utilisés pour construire la carte génétique de la chicorée industrielle

Marqueurs AFLP												SSR génomique		EST-SSR	SNP	
A1678	B1820	D1353	E1780	H2726	Jh	Ld	B1580	D2038	G1080	Im	Lj	GA647	GA075	est171	est40	bip23
A2147	B1966	D1489	E1835	H2962	Ji	Lh	B1641	D2215	G1226	Jc	Lk	GA355	GA336	est11	est18	bip28
A2217	B2006	D1577	ee117	H3003	Ka	Li	B1785	D2242	G725	Je1	Lm	ATC004	GA187	est8	cifr07	bip33
A2308	B2031	D1902	F1254	la	Kb	A1192	B1837	dd203	G782	Je3	Ln	GA565	GA5192	est46	feh1	bip37
A2317	B2078	D1931	F1370	lb	Kc	A1533	B1896	dd362	H1760	Jg	Lo	GA432	GA698	est5	pa5	bip41
A2335	bb161	D1963	F1436	lc	Ke	A2093	B2010	E1351	H1991	Jk		GA014	GA499	est7	pa14	bip43
A2454	bb254	D2005	F1541	lc1	Kh	A2134	B2096	E2023	H2507	Jl	AA T001	GA019	est41		pa21	bip44
aa68	C1089	D2350	F1557	ld	Kj	A2151	B2123	F1264	H2566	Kd		GA681	GA357	ms1	pa37a	bip70
B1130	C1301	D2403	F1937	lh1	Kk	A2153	B397	F1597	H2807	Kf		GA495	GT082	est15	pa37b	bip66
B1406	C1307	D400	G880	lj	Kl	A2267	bb260	F1748	lf	Kn		GA561	GA036	est1	pa39	bip49
B1435	C1497	D579	G936	lk	Km	A2484	bb264	F1797	lg	Kp		GT082	ciga23	est14	bip2e	bip59
B1439	C1516	D721	H1681	ll	Ko	A2532	C1220	F1821	lh	Kr		GA118	ciga24	ms10	bip02	bip57
B1585	C366	E1029	H1879	lb	Kq	aa101	C1389	F1904		La		ATC002	GA405		bip03	
B1599	C772	E1373	H2211	Je	Lb	B1039	C1504	F845	lj1	Lf		GA663	GA571		bip20	
B1708	D1069	E1518	H2602	Jf	Lc	B1137	C618	ff97	kl1	Lg		TGA001			bip21	
176												29		12	27	

IV.II. CONSTRUCTION DE LA CARTE GÉNÉTIQUE

Afin d'assigner chaque marqueur au sein d'un groupe de liaison, la valeur maximale de recombinaison a été fixée à 40 cM et le LOD score à un minimum de 7. Ces valeurs ont été déterminées suite aux observations faites sur l'influence de l'augmentation et de la diminution de ces deux critères (Claire-Noël Thienpont, ECAV-UCL).

Cette première étape a permis de former 9 groupes de liaison possédant un minimum de 18 marqueurs. Deux groupes de liaison supplémentaires possédant moins de cinq marqueurs ont également été formés. Ceux-ci n'ont pas été pris en considération pour la suite de l'analyse car ces marqueurs n'étaient pas considérés comme une base suffisante pour la construction d'un groupe de liaison. Le nombre de groupe de liaison correspond donc au caryotype de la chicorée (n=9).

La deuxième étape consistait à établir l'ordre des loci au sein de chaque LG. Afin de réduire le nombre de calculs à réaliser par le logiciel, les marqueurs hautement informatifs ont été sélectionnés. Ces marqueurs possédaient un génotype pour minimum 225 individus et étaient distants l'un de l'autre d'au minimum 5 cM. Seuls les marqueurs AFLP provenant de l'UCL et les marqueurs SSR provenant de la société Cosucra possédaient un génotype pour l'ensemble de la population SAMCHIC soit pour 247 individus. L'ensemble des ADN n'étant plus disponible, les marqueurs géniques de types SSR et SNP ont été utilisés pour le génotypage de 200 individus de la population de ségrégation.

Les critères de sélection des marqueurs hautement informatifs excluent donc ces marqueurs qui sont venus s'ajouter plus tard dans la construction de la carte.

108 marqueurs hautement informatifs ont été retenus et ordonnés au sein de leur groupe de liaison respectif (**figure 4.3** : marqueurs en gras). Nous avons obtenu au final une carte « framework » de 1086 cM composée de neuf groupes de liaison. La taille moyenne des LG était de 120,7 cM et la distance moyenne entre les marqueurs était de 10 cM.

Les marqueurs non-informatifs associés à un LG lors de la première étape ont ensuite été ajoutés à la carte « framework ». Au total, 104 marqueurs supplémentaires (**figure 4.3** : marqueurs soulignés) ont été positionnés, générant ainsi une carte composée de 219 marqueurs. Les distances génétiques pour chaque LG ont été calculées à partir de cette carte.

Un LOD score de 4 a ensuite été utilisé afin de positionner les 25 marqueurs restants. 18 marqueurs ont ainsi pu être placés sur la carte (**figure 4.3** : marqueurs en italiques). Ces derniers n'ont pas été pris en compte pour le calcul des distances. En effet, vu la valeur relativement faible du LOD score, la probabilité que l'assignation de ces marqueurs soit erronée est 1.000 fois plus élevée que pour les autres marqueurs. Il faut donc garder en tête que la position de ces marqueurs peut être inexacte. Seuls 7 marqueurs (3,2 %), tous provenant de l'AFLP, n'ont pas pu être cartographiés.

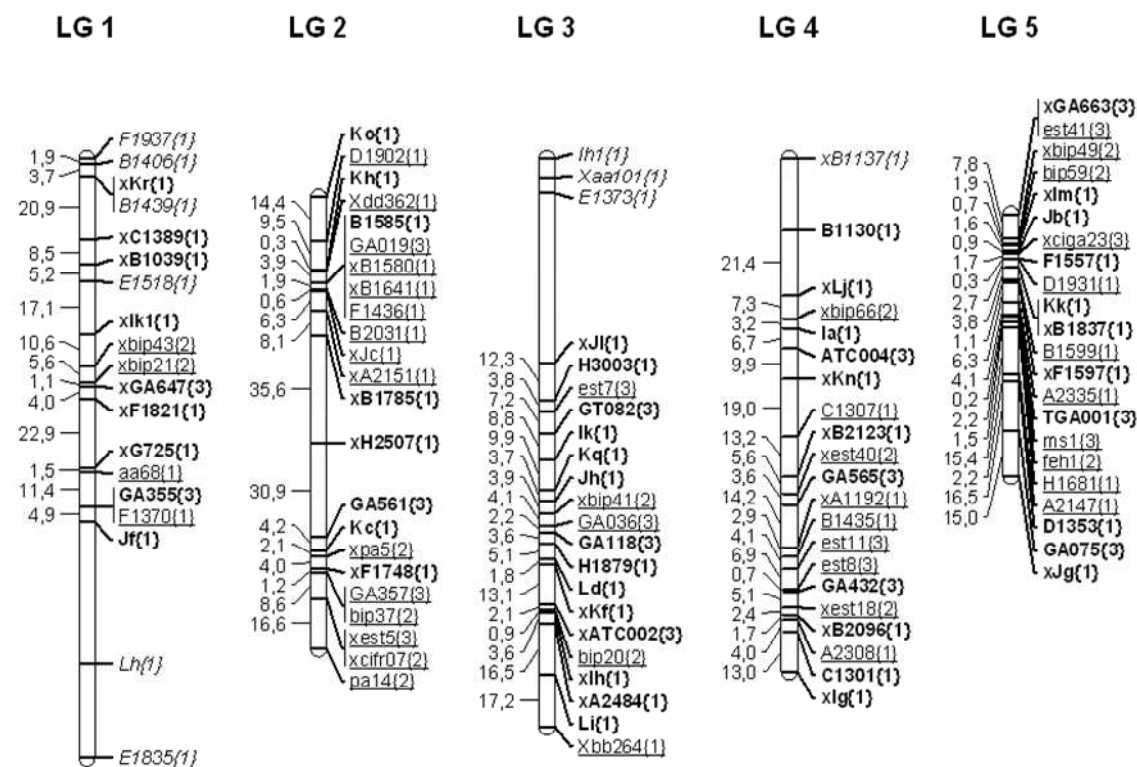


Figure 4.3 a : Représentation des neuf groupes de liaison (LG) de la carte génétique de la chicorée construits à partir de 244 marqueurs et de 247 individus de la population SAMCHIC. Les distances génétiques sont exprimées en centiMorgan (cM) Haldane. Gras : marqueurs positionnés sur les LG à un LOD de 7 et utilisés pour construire le framework. Souligné : marqueurs positionnés sur les LG à un LOD de 7 mais ne faisant pas partie du framework. Italique : marqueurs positionnés sur la carte à un LOD de 4. Ces marqueurs n'interviennent pas dans le calcul des distances génétiques. (1) marqueurs AFLP, (2) marqueurs SNP, (3) marqueurs SSR

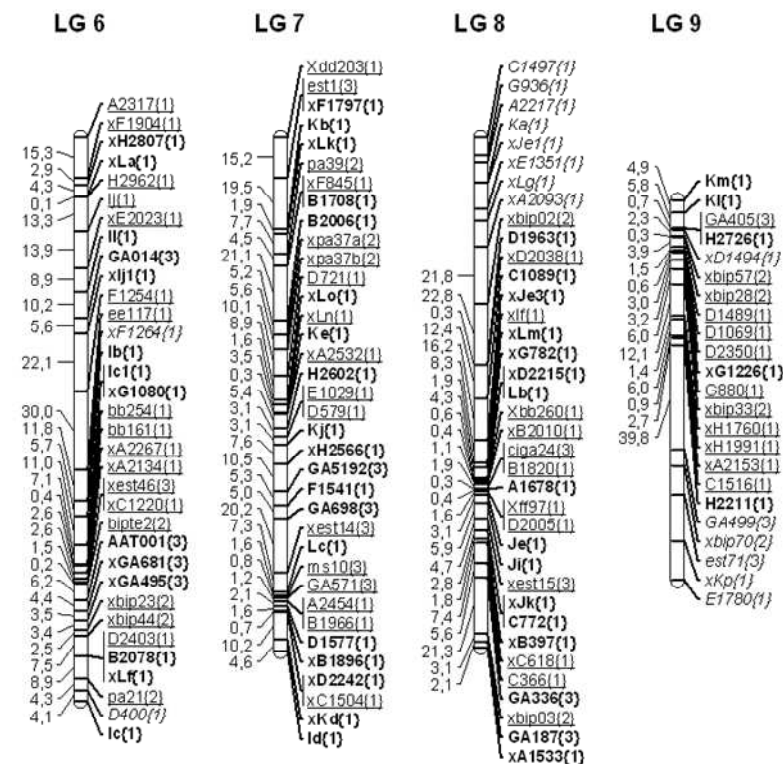


Figure 4.3 b: Représentation des neuf groupes de liaison (LG) de la carte génétique de la chicorée construits à partir de 244 marqueurs et de 247 individus de la population SAMCHIC. Les distances génétiques sont exprimées en centiMorgan (cM) Haldane. Gras : marqueurs positionnés sur les LG à un LOD de 7 et utilisés pour construire le framework. Souligné : marqueurs positionnés sur les LG à un LOD de 7 mais ne faisant pas partie du framework. Italique : marqueurs positionnés sur la carte à un LOD de 4. Ces marqueurs n'interviennent pas dans le calcul des distances génétiques. (1) marqueurs AFLP, (2) marqueurs SNP, (3) marqueurs SSR

IV.III. CARACTÉRISTIQUES DE LA CARTE GÉNÉTIQUE

La carte finale (**figure 4.3**) possède un total de 237 marqueurs répartis sur 9 groupes de liaison et recouvrant 1208 cM. La taille des LG varie entre 55,3 et 214,4 cM avec une moyenne de 132,6 cM. La distribution des intervalles entre les marqueurs suit une courbe exponentielle où la plus petite valeur (entre 0 et 1 cM) est la plus fréquente. La distance entre les marqueurs varie entre 0 et 35,6 cM avec une moyenne de 5,6 cM.

La sélection assistée par marqueur nécessite l'utilisation de marqueurs étroitement liés au locus cible, de préférence à une distance génétique de moins de 5 cM (Collard and Mackill 2008). En effet, l'utilisation de marqueurs flanquants ou intragéniques augmente considérablement la fiabilité de ces outils pour prédire le phénotype. Dans notre cas, si l'on considère la distance moyenne entre les marqueurs, n'importe quel locus se trouve à maximum 2,8 cM d'un marqueur. Cette carte génétique semble donc adaptée à la sélection assistée par marqueurs.

Au total, 152 marqueurs AFLP, 27 marqueurs SNP et 40 marqueurs SSR ont été positionnés sur la carte génétique à un LOD score de 7. La probabilité que ces marqueurs soient liés est 10^7 fois plus grande que l'hypothèse d'indépendance. Les groupes de liaison peuvent donc être considérés comme fiables.

Concernant les marqueurs anonymes de type AFLP, chaque LG possède en moyenne 17 marqueurs de ce type. Cela représente environ un marqueur tout les huit cM. Le rôle de ce type de technique est d'obtenir rapidement un grand nombre de marqueurs afin de construire une carte génétique dont le nombre de groupes de liaison correspond au nombre de chromosomes. L'objectif a été atteint puisque la chicorée possède 9 paires de chromosomes homologues.

Les marqueurs de type SNP et SSR sont représentés sur chaque groupe de liaison avec un minimum de quatre et un maximum de neuf marqueurs par LG. Cela représente une moyenne de 7 marqueurs par groupe de liaison, soit un marqueur tous les 18 cM. Ces marqueurs non-anonymes permettent d'amplifier spécifiquement un locus particulier dont on connaît la séquence et le type de polymorphisme. Ces marqueurs sont donc facilement utilisables pour le génotypage de chicorées d'intérêts agronomiques. De plus, dans le cas où aucun QTL

n'est positionné sur cette carte, ces marqueurs peuvent être utilisés pour génotyper un nouveau croisement et donc être localisés sur une nouvelle carte plus propice à l'analyse de traits quantitatifs.

Parmi les marqueurs SSR et SNP, 39 marqueurs ont été localisés au sein de séquences exprimées (EST). Il est donc possible de rechercher au sein d'espèces proches, les orthologues de ces séquences et donc de réaliser une cartographie comparée. Cela permet d'un point de vue fondamental de mieux comprendre l'évolution et l'organisation du génome. D'un point de vue plus appliqué, ce type d'analyse permet de savoir si le transfert de données génétiques est possible d'une espèce à l'autre.

IV.IV. INFLUENCE DE LA TAILLE DE LA POPULATION ET DU NOMBRE DE MARQUEURS

- Influence de la taille de la population

Plus une population de ségrégation est importante, plus il y a de chances d'observer des événements de recombinaison rares. Cela permet donc d'avoir une carte génétique plus dense. Par exemple, une population de ségrégation de 100 individus permettra la mise en évidence d'au minimum un événement de recombinaison ayant lieu une fois sur 100. Cela implique que la distance génétique de cette carte ne pourra être inférieure à environ 1 cM.

La taille de la population influence également la fiabilité de l'ordre des marqueurs. En effet, l'écart-type de la fréquence de recombinaison est inversement proportionnel au nombre d'individus (n) :

$$s_{\hat{r}} = \sqrt{\frac{1}{ni_r}}$$

Où S_r est l'écart-type de la fréquence de recombinaison, n est le nombre d'individus et i_r est la quantité d'informations qu'apporte chaque individu pour l'estimation de la liaison. Un n élevé donnera des écarts-types faibles. L'ordre des marqueurs sera donc plus fiable.

▪ Influence du nombre de marqueurs

Le nombre de marqueurs positionnés sur une carte peut également influencer le calcul du taux de recombinaison et donc la taille totale de la carte. En effet, si la densité des marqueurs est trop faible, la probabilité de ne pas observer les doubles crossing-over augmente. La **figure 4.4** représente un double crossing-over entre les marqueurs A/a et C/c. En absence du marqueur B/b, ces deux événements de recombinaison ne pourront être observés. La valeur de r sera donc sous estimée et la taille de la carte réduite. Au plus le nombre de marqueurs sera élevé, au plus la fréquence de recombinaison calculée sera proche de la valeur réelle.

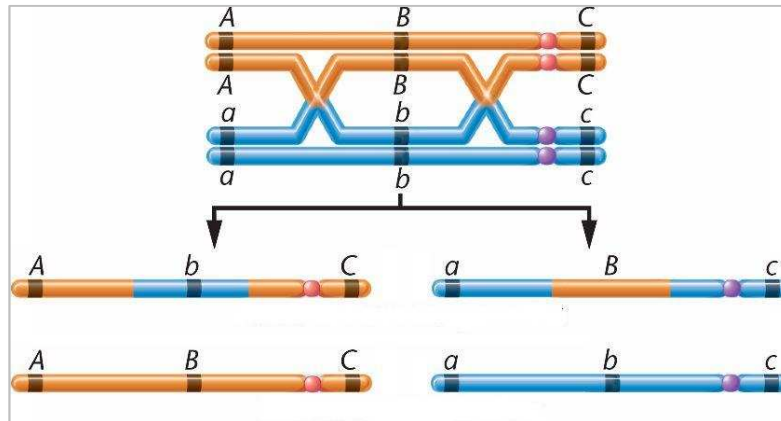


Figure 4.4 : Représentation d'un double crossing-over entre les locus Aa, Bb et Cc. Au final, la moitié des gamètes possède un chromosome recombinant. Le génotypage du locus Bb permet de mettre en évidence le double crossing-over. En cas d'absence de ce marqueur, les deux événements de recombinaison ne pourront pas être identifiés.

Le **tableau 4.3** reprend l'estimation du nombre de marqueurs nécessaires pour obtenir une proportion de recouvrement (P) du génome, en fonction de la taille du génome (k) et de l'espacement maximal (c), en cM, entre les marqueurs (Beckmann and Soller 1983). Pour ces données, on considère que le génome est circulaire. Afin de contrer les effets de fin de chromosome, ces valeurs doivent être multipliées par 1,2 pour $c=10$ et par 1,3 pour $c=20$.

Notre carte a une taille de 1.208 cM. Cela signifie que le nombre de marqueurs nécessaire pour avoir un recouvrement de 95% se situe aux alentours de 250 marqueurs ($178 \times 1.2 \times 1.2$) pour un génome de cette taille. Notre jeu de 244 marqueurs semble donc suffisant pour

obtenir cette proportion de recouvrement. Nous en rediscuterons dans le point consacré à la taille estimée du génome ci-dessous.

Tableau 4.3: Nombre de marqueurs nécessaires pour obtenir une proportion de recouvrement (P) du génome, en fonction de la taille du génome (k) et de l'espacement maximal (c), en cM, entre les marqueurs (Beckmann et Soller, 1983)

Proportion de recouvrement du génome	Taille du génome (k) et espacement (c), en cM					
	k = 1000		k = 2000		k = 3000	
	c = 10	c = 20	c = 10	c = 20	c = 10	c = 20
0,50	41	22	83	45	124	67
0,60	54	29	109	59	164	89
0,70	71	38	144	77	215	117
0,80	96	51	192	104	287	156
0,90	137	73	275	148	411	224
0,95	178	95	358	193	535	291

- Comparaison des différentes cartes génétiques de *C. intybus*

Le **tableau 4.4** reprend les différentes caractéristiques de chacune des cartes génétiques construites pour *C. intybus*. Les cartes K28K29, rubis118 et BR correspondent aux cartes génétiques utilisées pour construire la carte consensus publiée par Cadalen en 2010.

Plusieurs facteurs entrent en jeu afin d'expliquer la variabilité observée pour la taille totale ainsi que la densité des différentes cartes.

Dans un premier temps, la carte de De Simone a été construite à l'aide d'un croisement interspécifique. Cette démarche permet d'augmenter le polymorphisme au sein de la population de ségrégation. Cela pourrait expliquer le nombre important de marqueurs polymorphes dans le croisement. Ce nombre important de marqueurs permet de compenser la faible taille de la population. En effet, s'il existe un événement de recombinaison, il sera plus que probablement observé. Cet équilibre entre un nombre important de marqueurs et un faible nombre d'individus a permis d'obtenir une carte dont la densité de marqueurs est plus élevée que la moyenne.

Parmi les cartes réalisées à l'aide d'un croisement intraspécifique, deux d'entre elles (Van Stallen et BR) ont été construites sur base d'un croisement de deux parents appartenant à *C. intybus* var. *foliosum*. Le polymorphisme présent au sein de cette variété est faible (Van Stallen 2003). Cela pourrait donc expliquer le nombre peu élevé de marqueurs utilisés pour la construction des deux cartes. Dans le cas de la carte de Van Stallen, la faible taille obtenue pourrait être liée à ce facteur. En effet, si la population est constituée d'un grand nombre

Tableau 4.4 : Caractéristiques des 6 cartes génétiques de *C. intybus*.

	Van Stallen et al. 2003	De Simone et al. 1997	K28K2 (Cadalen et al. 2010)	Rubis 118 (Cadalen et al. 2010)	BR (Cadalen et al. 2010)	SAMCHIC	Moyenne
Croisement	intraspécifique	interspécifique	intraspécifique	intraspécifique	intraspécifique	intraspécifique	-
Espèce	<i>C. intybus</i> L. var foliosum	sylvestre x <i>C.</i> endiva	<i>C. intybus</i> L. var sativum	<i>C. intybus</i> L. var sativum	<i>C. intybus</i> L. var foliosum	<i>C. intybus</i> L. var sativum	-
Population de cartographie	565 individus F2	46 individus F1	177 individus F2	96 individus F2	145 individus F2	247 individus F2	213
Marqueurs cartographiés	126	371	274	282	178	237	245
Taille de la carte (cM)	609,6	1330	749,8	806,1	891,4	1208	
Distance moyenne entre les marqueurs (cM)	4,7	3,6	2,7	2,8	5	5,6	4

d'individus mais qu'il y a peu de marqueurs pour observer les recombinaisons, alors il est probable que la taille de la carte soit sous-estimée.

Les deux cartes les plus denses (K28K29 et Rubis118) correspondent à des croisements intraspécifiques. Pour ces deux cartes, la taille de la population n'est pas très importante mais le nombre de marqueurs est plus élevé que pour les autres cartes intraspécifiques. Il semble donc que ce soit ce dernier facteur qui a permis d'obtenir ces densités plus élevées par rapport aux autres cartes.

Dans notre cas, le nombre d'individus utilisés se situe au-dessus de la moyenne des différentes cartes et le nombre de marqueurs se situe légèrement en-dessous de la moyenne. La taille totale de la carte est parmi les plus grandes et l'écart moyen entre les marqueurs est le plus élevé. Deux facteurs pourraient expliquer ce phénomène. Tout d'abord, il se peut que la distance physique entre les marqueurs utilisés dans notre carte soit plus importante que pour les marqueurs des autres cartes. Ensuite, un taux de recombinaison plus élevé dans notre croisement pourrait expliquer cette densité plus faible. Afin d'augmenter la densité de notre carte, l'augmentation du nombre de marqueurs semble une approche appropriée.

IV.V. CALCUL DE LA TAILLE ESTIMÉE DU GÉNOME

La taille du génome est définie ici par la somme des tailles de groupes de liaison (en cM). Lorsque la carte n'est que partiellement saturée du fait d'un nombre de marqueurs insuffisant, cette somme est inférieure à celle que l'on obtiendrait si elle était saturée. L'estimation de la taille du génome permet donc de connaître la taille de la carte si celle-ci était saturée. Cette valeur peut servir à établir une relation entre distance de carte (cM) et distance physique (en nombre de paires de bases) (Lorieux 1994c).

Cette valeur a été calculée suivant l'expression de Fishmann et de Chakravarti (voir matériel et méthode ci-dessus). La méthode de Fishman ajoute pour chaque groupe de liaison, deux fois l'intervalle moyen entre les marqueurs. Cela représente les deux extrémités de chaque LG. La méthode 4 de Chakravarti multiplie la longueur de chaque LG par le facteur $(k+1)/(k-1)$ où k est le nombre de marqueurs

positionnés sur le groupe de liaison. Ces deux méthodes ont été utilisées dans l'article de Cadalen et al. en 2010.

Les deux calculs ont donné les mêmes résultats : la taille estimée du génome calculée à partir de la carte « framework » est de 1.290 cM. Cela signifie que la carte « framework » recouvre 84% du génome total. En faisant le calcul avec tous les marqueurs positionnés sur la carte avec un LOD score de 7, nous avons obtenu une taille estimée du génome de 1.318 cM. Notre carte recouvre donc au final 92% du génome total.

L'estimation du nombre de marqueurs nécessaires pour avoir un recouvrement du génome de 95% est de 250 (voir plus haut). Le calcul de la taille estimée du génome a été fait sur base des 219 marqueurs positionnés à un LOD score de 7. Une trentaine de marqueurs supplémentaires devrait donc permettre d'obtenir un pourcentage de recouvrement optimal.

IV.VI. CORRESPONDANCE cM-PAIRES DE BASE

La taille physique du génome est estimée chez *C. intybus* à 1.333 Mb (Koopman 2002). Si l'on considère la taille estimée du génome (1.315 cM), 1 cM sur notre carte représente 1 Mb. Cela signifie que les marqueurs sont en moyenne distants de 5,6 Mb.

Chaque carte possède sa propre correspondance. Un cM représente 0,99 Mb sur la carte de De Simone, 1,92 Mb sur la carte de Van Stallen et 1,53 Mb sur la carte de Cadalen. Cette valeur permet d'obtenir une idée de la distance physique séparant les marqueurs. Cela peut être intéressant pour la caractérisation des QTL ainsi que pour le clonage positionnel. Cependant, il ne faut pas oublier que cette correspondance, tout comme la taille, sont des valeurs intrinsèques à chaque carte en fonction de la population utilisée, du nombre de marqueurs, ... Ce ne sont en rien des valeurs absolues pour l'espèce étudiée, contrairement à la taille physique du génome.

IV.VII. FONCTIONS ASSOCIÉES AUX SÉQUENCES CARTOGRAPHIÉES

Sur la carte, 39 marqueurs se trouvent au sein de séquences exprimées. Le **tableau 4.5** reprend ces différentes séquences, leur numéro d'accèsion dans NCBI ainsi que les fonctions de séquences homologues qui leurs sont associées avec également leur numéro

d'accension. Parmi celles-ci, certaines sont liées à des fonctions d'intérêt pour la sélection de la chicorée :

- Est1(FL675136) : ce marqueur est associé à une séquence homologue de SKU5 provenant d'*A. thaliana*. SKU5 est une protéine impliquée dans la croissance directionnelle de la racine. Elle agirait via la modification des liaisons entre les composants structuraux de la paroi cellulaire (Sedbrook et al. 2002).
- Bip02 (EH672777) : la séquence associée à ce marqueur est un homologue du gène codant pour une protéine TIR-NBS-LRR (Toll/interleukin 1 receptor Nucleotide Binding Site Leucine-Rich Repeat). Cette protéine semble jouer un rôle dans la résistance de type I chez le tournesol face à *Plasmopara halstedii*, l'agent responsable du mildiou (Radwan et al. 2011).
- Bip66 (EH691923) : l'homologue de cette séquence est une Wall Associated Kinase (WAK). Cette protéine membranaire est impliquée dans l'expansion cellulaire, la morphogenèse mais également dans la réponse aux pathogènes. En effet, WAK 1 est un récepteur capable de lier des fragments pectiques produits lors de l'action de pectinases fongiques et bactériennes sur la paroi végétale (Decreux and Messiaen 2005).
- Bip59 (EH682796) : la recherche d'homologie a permis d'identifier un gène impliqué dans la floraison : GIGANTEA (GI). GI est une des protéines impliquées dans le contrôle de la floraison et plus particulièrement dans la perception de la photopériode (Sawa and Kay 2011).

Plusieurs marqueurs sont associés à des enzymes en relation directe ou indirecte avec le métabolisme de l'inuline, dont les invertases. Il existe deux groupes d'invertases (également appelées β -fructofuranosidases) : les invertases acides et les invertases neutres/alcalines. Les invertases acides sont divisées en deux catégories : les invertases pariétales et les invertases vacuolaires (Ji et al. 2005). Les enzymes impliquées dans le métabolisme de l'inuline semblent toutes provenir d'invertases acides. La 1-SST (Sucrose-1-Sucrose Fructosyltransferase) et la 1-FFT (Fructose-1-Fructose Fructosyltransferase) seraient issues de l'évolution d'une invertase vacuolaire (Ritsema et al. 2006) et les FEH (Fructan exohydrolase), d'invertases pariétales (Van den Ende et al. 2000).

Tableau 4.5 : Récapitulatif des marqueurs cartographiés associés à une séquence exprimée

Catégorie	Nom du marqueur	Nom de la séquence	Type de séquence	LG	Type de marqueur	Numéro d'accension GenBank	Description du Best Blastx	Fonction	Espèce	Accession du best Blastx	eValue	Positifs	Similitude
Métabolisme du carbone	leh1	leh1	ADNg	5	SNP	AJ242538	Fructan 1-xylohydrolase I	Dépolymérisation de l'inuline	Cichorium intybus	CAC19366.1	0.0	568/568	100%
Métabolisme du carbone	est11	REN002402 SCF	EST	4	SSR	FL675360	Glycosyltransferase, CAZy family GT8	Transfert de groupement glucosyl	Populus trichocarpa	XM_002329906.1	1.05E-47	67/74	90%
Métabolisme du carbone	est14	Cich_inty.EH683296	EST	7	SSR	EH683296	Brittle 1 protein-like	Transport de l'ADP-glucose	Solanum tuberosum	DQ207850.1	7.04E-36	77/87	88%
Métabolisme du carbone	est5	NOG10G12	EST	2	SSR	FL672278	1-SST	Initiation de la synthèse de l'inuline	Cichorium intybus	U81520.1	3.39E-41	47/59	79%
Métabolisme du carbone	est7	RBL009G06.GB	EST	3	SSR	FL673198	SNF1-related protein kinase	Réponses cellulaires diverses	Cucumis sativus	Y10036.1	1.44E-02	45/45	100%
Métabolisme du carbone	est8	CCIM8568.b1_P05.ab1	EST	4	SSR	EH696752	Trehalose-6-phosphate synthase	Synthèse du tréhalose	Ricinus communis	XM_002522209.1	9.39E-65	106/116	91%
Métabolisme du carbone	bip03	CCIM9953.B1_B18	EST	8	SNP	EH698135	Beta-fructofuranosidase	Invertase	Ricinus communis	XM_002531965.1	2.50E-156	246/265	92%
Métabolisme du carbone	ch07	ch07	ADNg	2	SNP	XJ082518	soluble acid beta-fructofuranosidase isoenzyme	Invertase	Daucus carota	X75333.1	1.65E-172	175/190	92%
Croissance	est1	REL008C07.GB	EST	7	SSR	FL675136	SKU5	Croissance directionnelle de la racine	Arabidopsis thaliana	NM_001203775.1	4.14E-75	174/182	79%
Croissance	est15	Cich_inty.EH691306	EST	8	SSR	EH691306	Alpha-expansin 3	Elongation de la paroi	Populus tremula	AY435101.1	8.84E-102	218/225	96%
Croissance	bip41	XC12_CONTIG792	EST	3	SNP	EL345653	coniferylalcohol 5-hydroxylase	Synthèse de la lignine	Centaurium erythraea	AAS92625.1	0.0	281/324	86%
Croissance	est40	XC12_CONTIG5886	EST	4	SNP	EH700985	lactoyglutathione lyase	Dégradation du méthylglyoxal	Ricinus communis	XM_002517560.1	4.90E-81	138/147	93%
Croissance	pa14	Contig173	EST	2	SNP	FL675436	S-adenosyl-L-methionine synthetase (SAM5)	Synthèse de la lignine	Chrysanthemum coronarium	F3461753.1	1.90E-116	205/213	96%
Facteur de transcription	bip49	AF101420.1	EST	5	SNP	EL353500	MDX3 box protein	Développement	Cichorium intybus	AF101420.1	1.07E-78	132/132	100%
Facteur de transcription	bip57	clfw12	EST	9	SNP	EH692696	transcription factor IIIA (TFIIIA)	Transcription d'ARN ribosomal	Arabidopsis thaliana	NM_105863.2	1.96E-71	48/58	82%
Facteur de transcription/Défence	est41	Contig2755	EST	5	SSR	EH692953	WRKY DNA-binding protein	Facteur de transcription impliqué dans la	Nicotiana tabacum	BAB61056.1	1.45E-139	301/498	60%
Facteur de transcription/Défence	est46	Contig5901	EST	6	SSR	EH687457	DREB3	Réponse de défense à la dessiccation	Solanum lycopersicum	NM_001247778.1	2.49E-64	106/114	92%
Facteur de transcription/Défence	bip02	CCIL10409.B1_A11	EST	8	SNP	EH672777	TIR-NBS-LRR resistance-like protein	Résistance au Mildew	Helianthus annuus	GU085221.1	1.27E-73	100/139	71%
Défense	bip33	CCIL10281.B1_B04	EST	9	SNP	E4672649	Elicitor inducible beta-1,3-glucanase NIEC-E76	Réponse aux pathogènes	Nicotiana tabacum	AB050628.1	1.20E-109	103/120	85%
Défense	bipe2	CCIL1586.B1_C13	EST	6	SNP	EH676972	barley mlo defense gene homolog6 (mlo6)	Réponse aux pathogènes	Zea mays	NM_001111700.1	1.39E-21	46/93	70%
Défense	est18	Cich_inty.EH705156	EST	4	SNP	EH705156	LEA-2 protein	Réponse au froid	Eupatorium chinense	F3357574.1	1.73E-58	125/151	82%
transduction du signal	bip37	XC12_CONTIG2587	EST	2	SNP	EH684149	ethylene-insensitive 2	Transduction du signal régulé par l'éthylène	Solanum lycopersicum	AA295507.2	0.0	445/717	62%
transduction du signal	bip66	CIC.EH691923	EST	4	SNP	EH691923	wall-associated kinase (WAK)	Signal dépendant de la pectine	Ricinus communis	XM_002516719.1	2.53E-116	206/243	84%
transduction du signal	bip70	Contig1503	EST	9	SNP	FL679098	CBL-interacting protein kinase	Réponse aux stress dépendant su calcium	Brassica napus	AAL37170.1	0.0	370/433	85%
Métabolisme des protéines	est71	Cich_inty.EH675186	EST	9	SSR	EH675186	Signal peptide peptidase-like 4 (SPPL4)	Endopeptidase	Arabidopsis thaliana	NM_001035953.1	1.20E-73	84/103	81%
Métabolisme des protéines	ms1	RKL007D02.GB	EST	5	SSR	FL677058	Cullin-1	Ubiquitination des protéines	Ricinus communis	XM_002516853.1	1.33E-97	169/171	98%
Métabolisme des protéines	pa21	Contig1536	EST	6	SNP	EL368712	ribosomal protein L5	Synthèse des protéines	Ricinus communis	XM_002524813.1	1.35E-80	159/166	95%
Métabolisme des protéines	pa5	Contig1092	EST	2	SNP	FL675412	elongation factor 1-alpha	Traduction	Cichorium intybus	AY378166.1	2.50E-120	210/227	92%
Divers	bip20	CCIM8748.B1_H02	EST	3	SNP	EH691568	Grave disease carrier protein	Transport mitochondrial	Ricinus communis	XM_002532617.1	1.47E-82	113/119	94%
Divers	bip21	CCIS4138.B1_D04	EST	1	SNP	EH705277	Rubisco subunit binding-protein alpha subunit	Photosynthèse	Trifolium pratense	AB236859.1	1.23E-58	126/136	94%
Divers	bip23	XC12_CONTIG2197	EST	6	SNP	EH682079	adenylosuccinate synthetase	Synthèse des acides nucléiques	Ricinus communis	XM_002526681.1	0.0	324/341	95%
Divers	bip59	Gladi	EST	5	SNP	EH682796	GLO GIGANTEA gene	Contrôle de la floraison	Lactuca saligna	HQ428558.1	3.28E-49	179/218	82%
Divers	pa37a/b	Contig942	EST	7	SNP	FL677179	NAD dependent epimerase	Epimerase	Trifolium pratense	AB236764.1	6.45E-63	114/120	95%
Divers	pa39	Contig32	EST	7	SNP	FL677172	met gene for cobalamine-independent methionine	Transfert de groupement methyl	Coleus blumei	Z49150.1	2.00E-82	138/154	95%
Pas de fonction connue	ms10	RBL005F06.GB	EST	7	SSR	FL672898	Clone LEFL1027DG07	Inconnue	Solanum lycopersicum	AK321709.1	1.80E-54	91/98	92%
Pas de fonction connue	bip28	XC12_CONTIG6915	EST	9	SNP	EH709102	ran-binding protein 3-like (LOC100082419)	Inconnue	Ornithorynchus anatinus	XM_003431152.1	1.53E+00	42/48	87%
Pas de fonction connue	bip43	CCIS6986.B1_L19	EST	1	SNP	EH707518	predicted protein	Inconnue	Populus trichocarpa	XM_002273892.1	1.10E-66	122/132	92%
Pas de fonction connue	bip44	CCIM1833.B1_B04	EST	6	SNP	EH689663	predicted protein	Inconnue	Populus trichocarpa	XM_002313404.1	4.83E-126	230/251	91%

Au total, cinq marqueurs positionnés sur la carte génétique sont associés à des invertases :

- FEH 1 (AJ242538) : cette séquence a été localisée sur la carte génétique au niveau du groupe de liaison 5. Cette enzyme, responsable de la dépolymérisation de l'inuline, est un candidat important pour l'amélioration des variétés de chicorée.
- Bip03 (EH698135) : cette séquence, associée à une invertase neutre/alkaline, a été localisée sur le LG 8 de la carte.
- Cifr07 (JQ082518) : localisée sur le LG2, cette séquence code pour une invertase vacuolaire (Dauchot 2007) dont l'homologue le plus proche est une β -fructofuranosidase vacuolaire (X75353.1), présente chez *Daucus carota* (eValue : $1,65 \times 10^{-172}$).
- Est5 (FL672278) : ce marqueur est localisé au même endroit que Cifr07 sur la carte génétique. Deux résultats de BLAST ont été retenus pour est5 car ils montraient des eValue similaires :
 - La même séquence que celle obtenue pour Cifr07 c'est-à-dire la β -fructofuranosidase vacuolaire présente chez *D. carota* (eValue : 4.5×10^{-41} , X75353.1)
 - La 1-SST de *C. intybus* (eValue : 3.4×10^{-39} , U81520.1)Vu la proximité des résultats, il est difficile de dire si Cifr07 et est5 font partie d'une seule et même séquence ou si est5 fait partie d'un autre gène issu également de l'évolution d'une invertase vacuolaire

Les fonctions décrites ci-dessus sont directement liées aux différents critères de sélection de la chicorée industrielle. Ces marqueurs peuvent donc être utilisés lors d'études d'association à grande échelle. De plus, puisque ce sont des marqueurs spécifiques de locus, ils peuvent être transférés dans d'autres variétés ou dans d'autres croisements en vue d'être utilisés pour la caractérisation de QTL ou encore pour la sélection assistée par marqueurs.

V. CONCLUSION

Plusieurs cartes génétiques de chicorée ont été publiées. Malheureusement, les caractéristiques intrinsèques des marqueurs cartographiés n'ont pas permis d'utiliser ces outils pour l'amélioration des variétés de chicorée étudiées. Nous avons dès lors entrepris de

contribuer à la construction d'une carte génétique en collaboration avec l'UCL et la société Cosucra.

Cette nouvelle carte génétique a été construite à l'aide d'une population F2 composée de 247 individus et de 237 marqueurs. La taille totale de la carte est de 1208 cM, ce qui représente une couverture estimée à 92% du génome total. Neuf groupes de liaison correspondant au nombre de paires de chromosomes chez *C. intybus* ont ainsi été formés.

La majorité des marqueurs utilisés sont de type AFLP. Cependant, 68 marqueurs co-dominants de type SNP et microsatellite ont été localisés sur la carte et sont potentiellement transférables. Parmi ceux-ci, 39 sont associés à une séquence exprimée pouvant avoir un intérêt pour la sélection de la chicorée, que ce soit dans la défense aux pathogènes, dans la montaison, dans la morphologie racinaire ou encore dans le métabolisme de l'inuline.

Afin de faire le lien avec le génome de la laitue et donc de bénéficier au mieux des ressources génétiques en plein développement de cette espèce proche, ces 39 séquences exprimées ont été utilisées pour la recherche d'orthologues afin de réaliser la première cartographie comparée entre les deux espèces. Notre carte peut donc servir de lien entre le génome de la chicorée et celui de la laitue. Cette ultime étape est détaillée dans le chapitre suivant.

Chapitre 5

Cartographie comparée entre

Lactuca et *C. intybus* var. *sativum*

Chapitre 5. CARTOGRAPHIE COMPARÉE ENTRE LACTUCA ET C. INTYBUS VAR. SATIVUM

I. BUT DE CE CHAPITRE

Lors du chapitre précédent, plusieurs dizaines de marqueurs associés à des séquences transcrites ont été positionnées sur la carte génétique de la chicorée. Dans ce présent chapitre, les orthologues de ces séquences ont été recherchés au sein de séquences positionnées sur le génome de la laitue via l'approche du Reciprocal best BLAST Hit (RBH). Une fois les paires d'orthologues mises en évidence, leur position au sein de leur génome respectif ont été comparées. Le but de cette démarche est de trouver des régions de synténie entre les deux espèces.

Cette cartographie comparée permet de bénéficier d'informations génétiques telles que les QTL ou les loci phénotypiques (également appelés marqueurs phénotypiques) localisés sur le génome de la laitue. Si ces informations génétiques sont présentes au sein de régions synténiques, alors il est possible que ces régions soient impliquées dans les mêmes processus chez la chicorée que ceux mis en évidence par les QTL et les loci phénotypiques chez la laitue. Cette démarche permet au final de sélectionner des régions du génome pouvant présenter un intérêt futur pour la chicorée.

Dans le présent chapitre, la mise en évidence des paires d'orthologues entre les deux espèces et la construction de la carte comparative ont été détaillées. Nous avons ensuite évalué la synténie et utilisé celle-ci afin de transposer de manière théorique chez la chicorée des données positionnées sur la carte de la laitue, telles que des QTL ou des loci phénotypiques.

II. INTRODUCTION

Avant l'émergence des méthodes de biologie moléculaire utilisées actuellement (séquençage, puces à ADN,...), la comparaison des génomes était réalisée principalement à l'aide de sondes RFLP dites hétérologues, c'est-à-dire des sondes d'ADN dérivants d'une espèce et utilisées pour détecter une séquence d'ADN similaire d'une autre espèce. L'analyse de la localisation de ces sondes au sein des génomes étudiés permettait de définir des régions de synténie (Moore et al. 1995). A ce jour, la disponibilité de nombreuses banques EST, de cartes physiques et de génomes entièrement séquencés permet de réaliser la comparaison de génomes via des analyses bioinformatiques. Ces analyses ont pour but de mettre en évidence des séquences orthologues entre les espèces étudiées. Les séquences orthologues sont des séquences homologues dérivées d'un événement de spéciation à partir d'une seule séquence ancestrale commune. De manière plus pragmatique, si deux séquences provenant de deux génomes différents se détectent réciproquement comme meilleur résultat de BLAST, alors ceux-ci sont définis comme orthologues (Moreno-Hagelsieb and Latimer 2008). La mise en évidence entre deux espèces, via la recherche de similitudes, de paires d'orthologues peut ainsi être utilisée pour la cartographie comparée.

La cartographie comparée est définie comme l'analyse de la localisation de séquences orthologues sur le génome d'espèces différentes. Elle permet d'identifier des régions synténiques, c'est-à-dire des régions du génome dont l'organisation est conservée entre les espèces étudiées.

Il existe deux niveaux de synténie : la macrosynténie et la microsynténie. Le degré de synténie dépend des outils utilisés lors de la cartographie comparée (Altman et al. 2011). Des méthodes telles que le chromosome painting, les analyses paires ou encore la cartographie génétique limitent le niveau d'analyse à la macrosynténie. La macrosynténie met en évidence uniquement la conservation de morceaux entiers de chromosomes ou de groupes de liaison. On parle de microsynténie lorsque la conservation est observée au niveau des séquences nucléotidiques. Alors que la macrosynténie peut être très élevée entre deux espèces étudiées, l'accumulation de mutations telles que des substitutions, des

insertions, des délétions, des duplications ou encore des réarrangements locaux peut détériorer la microsynténie.

L'intérêt de la cartographie comparée réside dans le fait que les résultats observés permettent d'émettre des hypothèses sur l'histoire évolutive du génome des espèces comparées. Cela donne en effet une idée des événements de recombinaison (fusion, fission, translocation, inversion, délétion, duplication) qui ont pu se passer depuis la divergence de deux espèces.

Chez les Asteraceae, très peu de choses ont été étudiées à ce sujet. Concernant l'origine de cette famille, on estime qu'elle est apparue entre le Miocène (20 Mya) et le Crétacé (100 Mya). Les tribus étaient présentes à la fin de l'Oligocène (22-35 Mya) et les genres à la fin du Miocène (5 Mya) (Funk et al. 2005). Concernant l'évolution des génomes, un événement de duplication totale a été identifié avant la séparation en tribus. Deux autres polyploïdisations ont eu lieu de manière indépendante à la base des tribus Heliantheae et Mutisieae (Barker et al. 2008) (chapitre 1 **figure 1.16**). Concernant la chicorée, il semblerait donc que le seul événement de polyploïdisation se soit déroulé bien avant l'apparition des Cichorioideae.

Une cartographie comparée entre deux espèces de la tribu Heliantheae, *Helianthus annuus* et *H. petiolaris* a permis de montrer un taux de réarrangement chromosomique très élevé (5,5 - 7,3 réarrangements par millions d'années) (Timme et al. 2007). Cela peut s'expliquer par l'événement de polyploïdisation survenu à la base de la tribu et qui est généralement suivi d'un grand nombre de réarrangements chromosomiques avec pour conséquence de diminuer le nombre de chromosomes (Salse 2012). A ce jour, cette étude de synténie est la seule au sein de la famille des Asteraceae.

D'un point de vue plus pratique, une connaissance approfondie des régions conservées permet également de transférer des informations génétiques d'une espèce à l'autre pour les zones synténiques. Par exemple, une cartographie comparée entre le génome du riz et du blé a été utilisée afin d'identifier des gènes candidats présents sur le génome du riz pouvant être utilisés pour caractériser un QTL lié à la viscosité des pentoses chez le blé (Quraishi et al. 2009).

Dans le cadre de ce travail, nous avons réalisé une cartographie comparée entre la chicorée industrielle et la laitue. En effet, la laitue

étant une espèce économiquement importante, les ressources génétiques développées au sein du Compositae Genome Project sont considérables. La carte physique de la laitue peut être consultée en ligne. Celle-ci est constituée de plus de 51.000 contigs composés de séquences transcrites (Koozik et al. 2011).

De plus, quatre études ont permis de co-localiser des loci phénotypiques et des QTL avec des EST localisées sur la carte physique de la laitue. Plusieurs loci phénotypiques et QTL associés à la résistance au mildiou, à l'anthracnose, aux pucerons des racines, au champignon *Plasmopora lactucea-radiciis*, à la maladie dite de la racine liégeuse ainsi qu'à la réponse hypersensible ont été positionnés sur une carte reprenant uniquement des EST associées à la résistance aux pathogènes (McHale et al. 2009). D'autres QTL et loci phénotypiques associés à la longévité de la graine suivant les conditions de stockage (Schwember and Bradford 2010), à la résistance au champignon *Verticillium dahlia* (Hayes et al. 2011) ou encore à la sensibilité à la triforine (un inhibiteur de la synthèse des stéroïdes présent dans les fongicides) (Simko et al. 2011) ont également été localisés à proximité de séquences EST présentes sur la carte physique de la laitue.

L'analyse de la synténie entre la carte génétique de la chicorée et la carte physique de la laitue permet de voir si ces loci se trouvent dans des régions conservées entre les deux espèces. Si c'est le cas, les séquences de chicorée localisées dans ces régions synténiques deviennent des gènes candidats positionnels pour l'étude des processus impliqués. Cette démarche permet ainsi de bénéficier des outils développés chez la laitue afin d'identifier et d'analyser des marqueurs d'intérêt chez la chicorée.

La recherche de paires d'orthologues entre les séquences positionnées sur la carte génétique de la chicorée et les séquences qui constituent les cartes génétiques et physiques de la laitue a tout d'abord été réalisée. Ensuite, la carte comparative a été construite afin d'identifier les zones de synténie entre les deux espèces. Enfin, nous avons recherché dans la littérature des loci phénotypiques ainsi que des QTL localisés chez la laitue dans ces zones de synténie afin de mettre en évidence des régions d'intérêts potentiels chez la chicorée.

III. MATÉRIEL ET MÉTHODES

III.I. BANQUES DE DONNÉES

- **Banque Lactuca JMR** : contient 435 EST cartographiées sur les cartes génétiques JMR 1, 2 et 3 provenant des espèces *Lactuca sativa* et *L. seriola* (CGP 2012). Chacune des EST a été extraite manuellement des trois cartes génétiques de laitue disponibles en ligne (étape réalisée par N. Dauchot) afin de former une banque de données locale à l'aide du programme Seqtools (<http://www.seqtools.dk>) (Rasmussen 2002-2009).
- **Banque Lactuca CLS** : contient 47.303 EST provenant de *L. sativa* dont une partie a été positionnée sur la carte physique de la laitue (CGP 2012). Le fichier contenant toutes les séquences est disponible sur le site et a été utilisé pour réaliser une banque de données locale à l'aide du programme Seqtools.

III.II. RECHERCHE DE SÉQUENCES ORTHOLOGUES ENTRE LA LAITUE ET LA CHICORÉE

Les séquences positionnées sur la carte de la chicorée ont été blastées (tBlastx) contre les banques Lactuca JMR et CLS. Cette étape a été réalisée à l'aide du programme Seqtools.

Les paires d'orthologues potentiels ont été sectionnées suivant deux critères : soit un RBH a été mis en évidence entre les deux séquences, soit le best BLAST hit obtenu possédait une e-Value inférieure à e^{-30} .

III.III. REPRÉSENTATION DES PAIRES D'ORTHOLOGUES SUR UNE CARTE COMPARATIVE

Seules les séquences ayant un orthologue potentiel chez la laitue ont été laissées sur la carte génétique de la chicorée. Concernant la carte de la laitue, la position des séquences est indiquée en pb. La position de séquences provenant de la carte génétique a été calculée proportionnellement entre le LG et le chromosome correspondant. La conception graphique de la carte a été réalisée à l'aide du programme MapChart (Voorrips 2002).

IV. RÉSULTATS ET DISCUSSION

IV.1. LA CARTOGRAPHIE CHEZ LA LAITUE

IV.1.1. Cartes génétiques

Une dizaine de cartes génétiques sont disponibles sur le site du CGP (<http://compgenomics.ucdavis.edu/>) pour la laitue. Sept cartes composées de marqueurs AFLP, RFLP, SSR et RAPD sont disponibles. A partir de celles-ci, une carte consensus de 2.744 marqueurs a été construite (Truco et al. 2007). De plus, trois cartes composées de plus de 435 EST dont les séquences sont disponibles ont également été mises en ligne. Enfin, une carte génétique construite à l'aide de 294 gènes candidats associés à la résistance aux pathogènes est également publiée et disponible sur le site (McHale et al. 2009). Le but étant la recherche de séquences orthologues à la chicorée, seules les cartes génétiques construites à l'aide de séquences EST ont été utilisées. Au total, 435 séquences EST localisées sur ces cartes ont été rassemblées au sein d'un seul fichier pour nos analyses (Banque *Lactuca* JMR).

IV.1.2. Carte physique

Dans le but de séquencer entièrement le génome de la laitue, une carte physique est en cours de construction (Koozik et al. 2011). A ce jour, environ 50.000 contigs constitués de séquences EST recouvrent environ 52 Mb. Les espaces entre les gènes ont été en partie cartographiés et pour l'instant, la carte physique représente 1,1 Gb soit un peu moins de la moitié du génome (2,7 Gb).

Une banque de données comportant la totalité des séquences EST positionnées sur cette carte n'est pas disponible. Cependant, il s'avère qu'une partie des EST provenant de la banque « CLS » de *L. sativa* (CGP 2012) est positionnée sur la carte physique. Cette banque comportant 47.303 séquences est disponible et a donc pu être utilisée pour la recherche d'orthologie (Banque *Lactuca* CLS).

Tableau 5.1 : Descriptif des 27 paires d'orthologues potentiels utilisés pour la cartographie comparée entre la laitue et la chicorée

Nom du marqueur	Numero d'accession	Position sur la carte de la chicorée	Nom de l'orthologue chez la laitue	eValue	Position sur la carte de la laitue	Type de carte chez la laitue	Methode d'obtention de l'orthologue	Nom de l'orthologue sur la carte comparative
bip21	EH705277	LG1	QG_CA_CONTIG282	9E-52	LG1	génétique	RBH	ortho.bip21
bip43	EH707518	LG1	QG_CA_CONTIG843	1E-127	LG1	génétique	RBH	ortho.bip43
bip37	EH684149	LG2	CLSM5414.B1_K09	1E-161	LG2	physique	RBH	ortho.bip37
est7	FL673198	LG3	CLSM18235.B1_F23	3E-07	LG8	physique	RBH	ortho.est7
bip20	EH691568	LG3	QG_CA_CONTIG2170	1E-122	LG3	génétique	RBH	ortho.bip20
bip41	EL345653	LG3	CLSM13971.B1_F13	1E-174	LG3	physique	RBH	ortho.bip41
est8	EH696752	LG4	CLSZ507.B1_F07.A	1E-106	LG4	physique	RBH	ortho.est8
est40	EH700985	LG4	CLSS11853.B1_J12	1E-122	LG4	physique	RBH	ortho.est40
ms1	FL677058	LG5	CLSY5195.B1_E04	0	LG5	physique	RBH	ortho.ms1
bip49	EL353500	LG5	CLSX9172.B1_H14	1E-157	LG6	physique	RBH	ortho.bip49
bip59	EH682796	LG5	CLSM19805.B1_I08	1E-127	LG9	physique	RBH	ortho.bip59
bip23	EH682079	LG6	CLSM2233.B1_B08	0	LG9	physique	RBH	ortho.bip23
bip44	EH689663	LG6	QG_CA_CONTIG1246	1E-60	LG1	génétique	RBH	ortho.bip44
Pa39	FL677172	LG7	CLSS10285.B1_J04	0	LG6	physique	RBH	ortho.Pa39
bip28	EH709102	LG9	CLSS3708.B1_G16	6E-47	LG9	physique	RBH	ortho.bip28
bip70	FL679098	LG9	CLSM2261.B1_J14	0	LG7	physique	RBH	ortho.bip70
bip57	EH692696	LG9	CLSM11354.B2_C08	1E-141	LG9	physique	RBH	ortho.bip57
Pa5	FL675412	LG2	CLSY2244.B1_H10	1E-118	LG2	physique	Best Blast Hit	ortho.Pa5
cifr07	JQ082518	LG2	CLSM14035.B1_E06	4E-80	LG5	physique	Best Blast Hit	ortho.cifr07
est11	FL675360	LG4	CLSY2664.B1_P18	6E-30	LG9	physique	Best Blast Hit	ortho.est11
bip66	EH691923	LG4	CLSM9490.B1_C22	1E-100	LG8	physique	Best Blast Hit	ortho.bip66
est41	EH692953	LG5	CLX_S3_CONTIG1123	1E-47	LG4	génétique	Best Blast Hit	ortho.est41
feh1	AJ242538	LG5	CLSM14035.B1_E06	1E-180	LG5	physique	Best Blast Hit	ortho.feh1
bipte2	EH676972	LG6	CLSM4674.B1_C17	6E-56	LG2	physique	Best Blast Hit	ortho.bipte2
bip02	EH672777	LG8	CLSX3769.B1_B08	9E-63	LG1	physique	Best Blast Hit	ortho.bip02
bip03	EH698135	LG8	QG_CA_CONTIG3540	4E-76	LG8	génétique	Best Blast Hit	ortho.bip03
bip33	EH672649	LG9	CLSM13713.B1_A22	1E-102	LG7	physique	Best Blast Hit	ortho.bip33

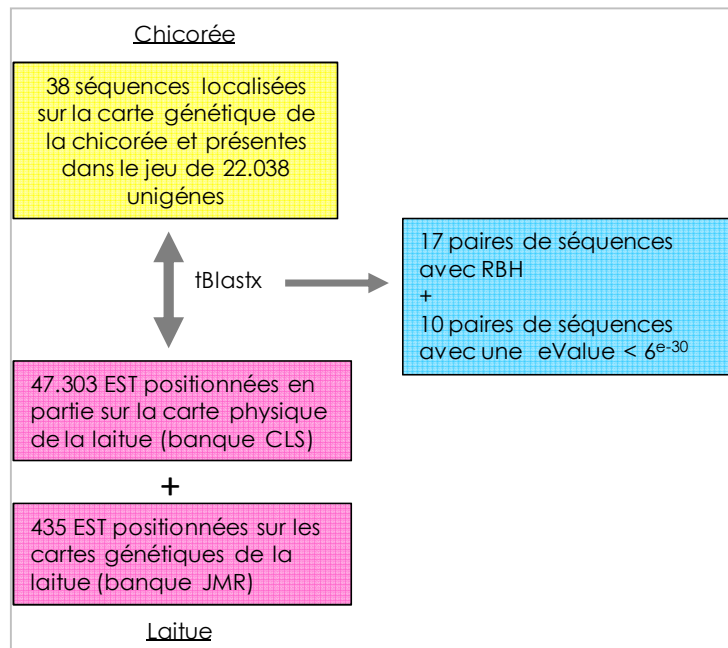


Figure 5.1 : Stratégie utilisée pour mettre en évidence des paires d'orthologues entre les séquences de chicorée et de laitue.

IV.II. RECHERCHE DE PAIRES D'ORTHOLOGUES

IV.II.1. Stratégie et résultats

Parmi les marqueurs géniques cartographiés dans ce travail, 12 marqueurs SNP sont situés au sein d'orthologues potentiels d'EST de laitue localisés sur les cartes génétiques. Depuis 2011, la carte physique de la laitue peut être consultée dans le cadre du Compositea Genome Project. Pour faire le lien avec cette carte, l'ensemble des séquences positionnées sur la carte de la chicorée, c'est-à-dire toutes les séquences associées aux marqueurs SNP et SSR géniques, a été utilisé pour la recherche d'orthologues chez la laitue. Le **tableau 4.5** du chapitre 4 reprend les informations concernant ces 38 séquences avec leur numéro d'accèsion ainsi que la fonction qui leur est associée.

L'ensemble des séquences positionnées sur la carte physique de la laitue ne peut être rassemblé dans un fichier afin de réaliser des analyses BLAST. Afin de contourner ce problème technique, la recherche d'orthologie a été réalisée à partir de la banque EST de *L. sativa* nommée CLS dont une partie se trouve sur la carte physique. La banque comportant les 435 EST positionnées sur les cartes génétiques de la laitue (Banque JMR) a également été utilisée pour la recherche d'orthologie (**Figure 5.1**).

Les 38 séquences de chicorée ont été blastées (tBlastx) contre les banques CLS et JMR de laitue. Les best BLAST Hit obtenus chez la laitue ont été blastés (tBlastx) contre la banque de 22.038 unigènes de chicorée (voir chapitre 2 point II.II.3.1) Au final, 17 séquences de chicorée possédaient un RBH avec une séquence positionnée sur une des cartes de la laitue. De plus, dix séquences de chicorée possédaient un best BLAST Hit dont la e-Value était de maximum $6e^{-30}$ avec une séquence de laitue cartographiée (**Figure 5.1**). Au final, 27 séquences de chicorée possédaient un orthologue potentiel sur la carte physique (21) ou sur la carte génétique (6) de la laitue (**tableau 5.1**).

IV.II.2. Fiabilité des résultats

IV.II.2.1. e-Value seuil

Lors de l'analyse BLAST des séquences, une e-Value est calculée afin d'estimer la fiabilité des résultats. L'e-Value représente la probabilité de

trouver par hasard un résultat d'alignement identique ou plus grand que l'alignement observé. Cette valeur prend en compte la taille de la banque de données ainsi que la composition de la séquence cible (Dear 2007). Plus les séquences sont semblables, plus la e-Value est faible.

La valeur seuil de l'e-Value obtenue lors du BLAST n'est pas facile à définir. En effet, il n'y a pas de règle précise pour ce critère. Cependant, si l'on regarde dans la littérature, plusieurs études de cartographie comparée au sein d'une même famille ont utilisé des e-Value seuils plus élevées et donc moins restrictives que dans notre analyse (e-Value de e^{-30}) :

- Cartographie comparée au sein des légumineuses (McConnell et al. 2010) : e-Value seuil de $1e^{-20}$
- Cartographie comparée au sein des Brassicaceae (Burrell et al. 2011) : e-Value seuil de $1e^{-07}$
- Cartographie comparée au sein des Poaceae (Yu et al. 2004) : e-Value seuil de $1e^{-10}$
- Cartographie comparée entre *A. thaliana* et deux Asteraceae (le tournesol et la laitue) (Timms et al. 2006) : e-Value seuil de $1e^{-25}$

Si le seuil de l'e-Value est trop élevé, la probabilité d'obtenir des alignements par chance augmente : des séquences qui semblent similaires sont en fait obtenues uniquement par hasard. Dans notre cas, la probabilité que nos résultats soient dus au hasard était de maximum $1/e^{30}$. Cela n'implique pas que les séquences sont des orthologues, ni même des homologues. Cela signifie que la similarité des séquences est telle que la chance que ce résultat soit dû au hasard est extrêmement faible.

Vu la valeur très faible de notre e-Value seuil, les paires de séquences obtenues par tBlastx ont été considérées comme fortement similaires. Cette similarité étant maximale pour ces paires de séquences par rapport aux autres EST de notre jeu de données, nous sommes partis de l'hypothèse que ces séquences étaient orthologues et non paralogues.

IV.II.2.2. Couverture du génome total par les banques utilisées

La recherche d'orthologues nécessite en théorie l'accès aux génomes complets des espèces étudiées, ce qui n'est pas encore possible pour la comparaison entre la chicorée et la laitue.

La taille des génomes de la chicorée et de la laitue a été estimée par mesure de la quantité d'ADN total (2C) par cytométrie de flux. Cette valeur est estimée à 6,044 pg pour la laitue et à 2,721 pg pour la chicorée (Koopman 2002). Un picogramme d'ADN correspond environ à 980 Mpb (Bennett et al. 2000). Cela signifie que la taille estimée du génome (1C) de la chicorée est de 1.333 Mpb et celui de la laitue, de 2.961 Mpb.

Pour la chicorée, c'est à ce jour la seule estimation disponible. Concernant la laitue, l'addition de la taille des différents chromosomes sur la carte physique donne une valeur de 2.625 Mb soit un peu moins que ce qui était estimé par Koopman en 2002. Puisque c'est à partir de cette carte physique que la cartographie comparée a été réalisée, nous gardons cette valeur (2.625 Mb) pour la taille du génome chez la laitue.

La recherche d'orthologie a été faite à partir d'une banque de 22.038 unigènes de chicorée représentant un total de 18,7 Mb. Cette banque de données représente donc approximativement 1,4% du génome total. Pour la laitue, les 47.303 EST utilisées représentent environ 24.000 unigènes pour un total de 20 Mb, ce qui représente 0,76% du génome total.

Le pourcentage de régions codantes au sein d'un génome est très variable. On estime que celui-ci représente, chez les angiospermes, moins de 10% du génome total (4% chez *A. thaliana* et 7.5% chez le maïs) (AGI 2000; Messing et al. 2004). Si l'on estime de manière grossière que 10% du génome est codant, alors nos données représentent 14% et 7,6% des régions codantes, respectivement chez la chicorée et chez la laitue. Il est évident que ce pourcentage est tout à fait hypothétique, mais cela nous permet de nous rendre compte de l'ordre de grandeur de nos données. En l'absence de séquençage complet de la chicorée, nous parlerons donc de paires d'orthologues potentiels.

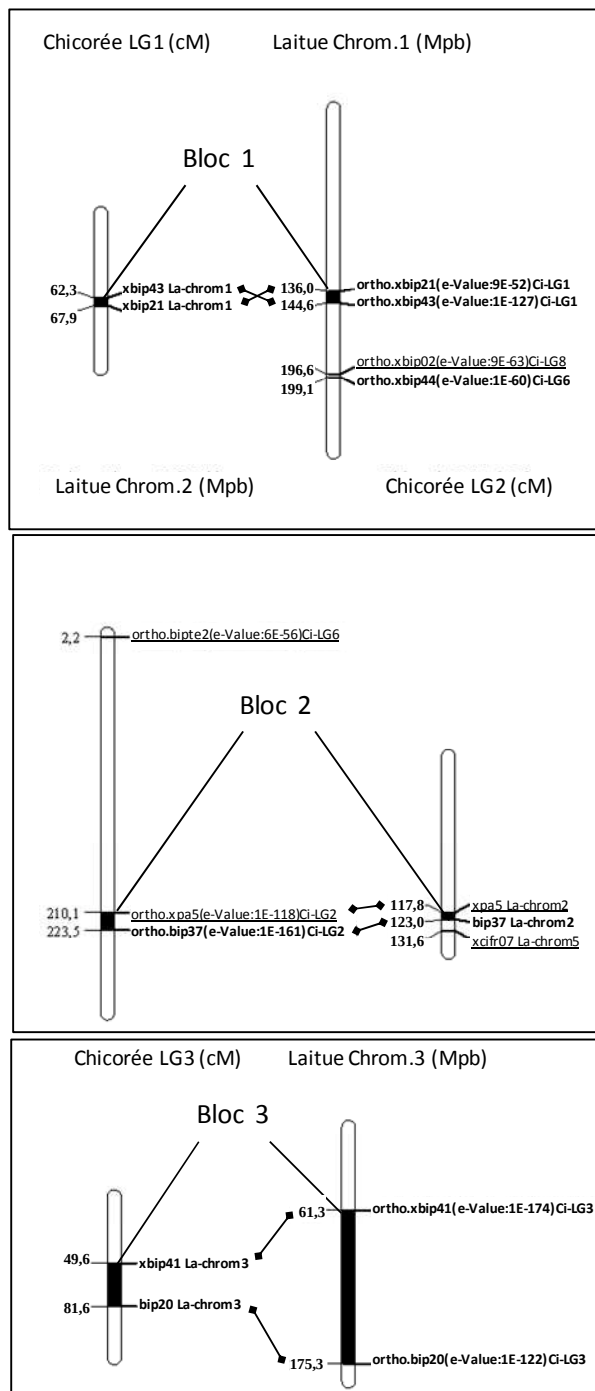


Figure 5.2.a : Cartographie comparée entre la chicorée industrielle et la laitue à l'aide de 26 paires d'orthologues potentiels. Sur chaque groupe de liaison de la chicorée (LG) sont indiqués : le nom du marqueur, sa position en cM Haldane ainsi que le nom du chromosome sur lequel se trouve l'orthologue potentiel chez la laitue. Sur chaque chromosome (Chrom.) de la laitue sont indiqués : le nom donné à la séquence, sa position en millions de paires de base, la eValue obtenue et le groupe de liaison sur lequel se trouve l'orthologue potentiel chez la chicorée

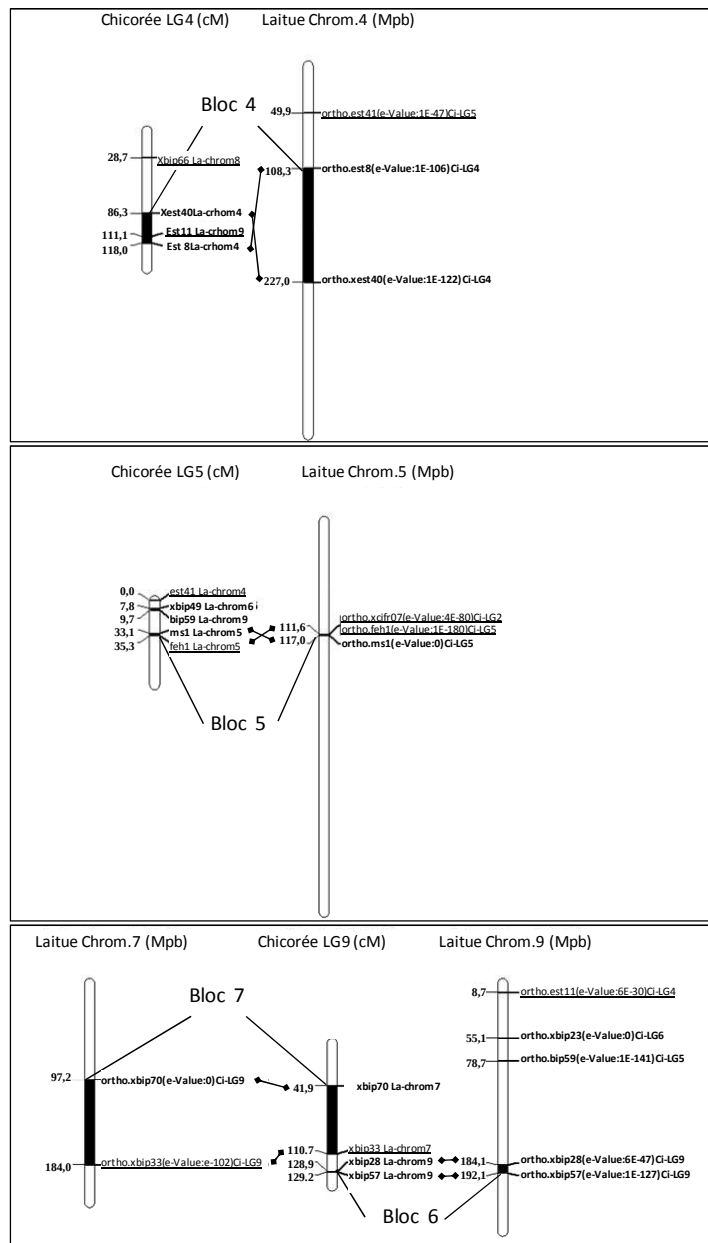


Figure 5.2.b : Cartographie comparée entre la chicorée industrielle et la laitue à l'aide de 26 paires d'orthologues potentiels. Sur chaque groupe de liaison de la chicorée (LG) sont indiqués : le nom du marqueur, sa position en cM Haldane ainsi que le nom du chromosome sur lequel se trouve l'orthologue potentiel chez la laitue. Sur chaque chromosome (Chrom.) de la laitue sont indiqués : le nom donné à la séquence, sa position en millions de paires de base, la eValue obtenue et le groupe de liaison sur lequel se trouve l'orthologue potentiel chez la chicorée

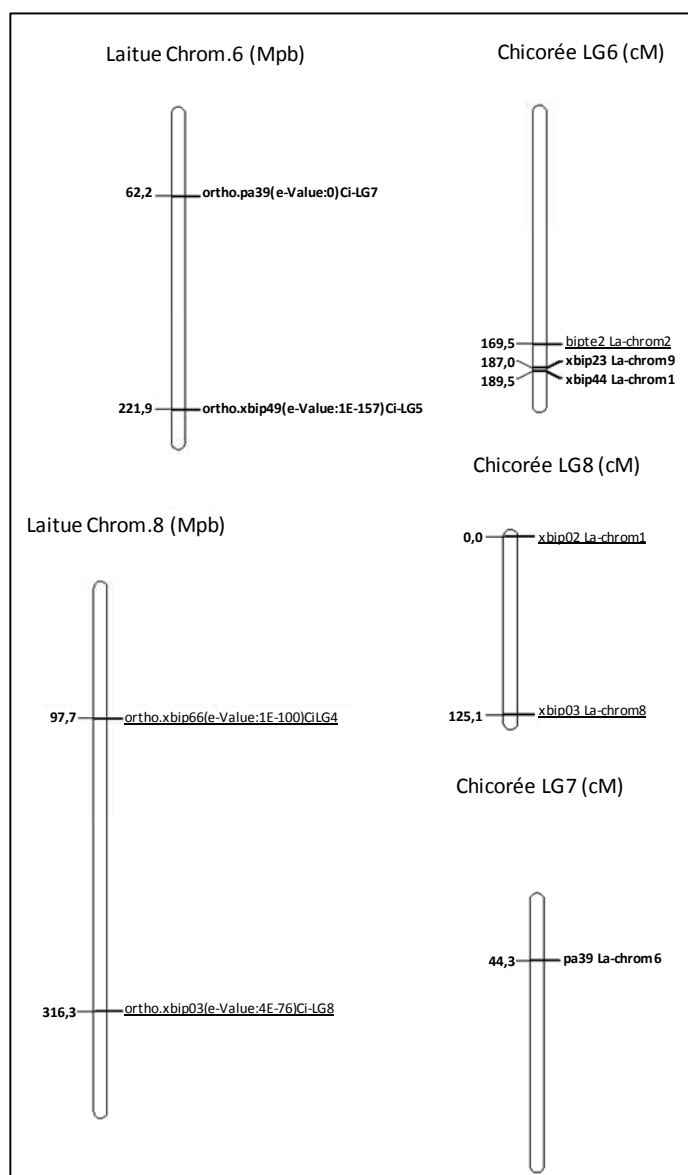


Figure 5.2.c : Cartographie comparée entre la chicorée industrielle et la laitue à l'aide de 26 paires d'orthologues potentiels. Sur chaque groupe de liaison de la chicorée (LG) sont indiqués : le nom du marqueur, sa position en cM Haldane ainsi que le nom du chromosome sur lequel se trouve l'orthologue potentiel chez la laitue. Sur chaque chromosome (Chrom.) de la laitue sont indiqués : le nom donné à la séquence, sa position en millions de paires de base, la eValue obtenue et le groupe de liaison sur lequel se trouve l'orthologue potentiel chez la chicorée

IV.III. ANALYSE DE LA SYNTÉNIE

IV.III.1. Construction de la carte comparative

La cartographie comparée a été réalisée à l'aide des 27 paires d'orthologues potentiels (**tableau 5.1**) entre la chicorée industrielle et la laitue. La **figure 5.2** représente la carte comparative pour ces paires d'orthologues potentiels. Les groupes de liaison de la chicorée sont en cM et les chromosomes de la laitue sont en millions de paires de base (Mpb).

Six orthologues proviennent de la carte génétique de la laitue. Afin de positionner ces séquences sur les chromosomes, leur localisation présumée en pb a été calculée à l'aide d'une règle de trois. Cela suppose que nous partons de l'hypothèse que les numéros de LG correspondent aux numéros de chromosome entre la carte génétique et la carte physique de la laitue. Nous n'avons pas eu confirmation de cette hypothèse par les auteurs de la carte physique de la laitue.

Afin d'approfondir cette question, un RBH a été réalisé entre la banque *Lactuca* JMR (EST positionnées sur la carte génétique) et la banque *Lactuca* CLS (EST positionnées en partie sur la carte physique). Pour 72.5% des RBH obtenus (127/175), le numéro du LG et le numéro du chromosome correspondaient. Cette analyse n'ayant pas été réalisée avec la totalité des séquences de la carte physique (non disponible), il est difficile de conclure sur la correspondance entre les deux types de carte. Nous avons donc considéré, comme nos résultats l'indiquent, qu'en majorité, les séquences positionnées sur un LG de la carte génétique se trouvent sur le chromosome correspondant sur la carte physique. Il faut cependant rester vigilant quant à la possibilité que ça ne soit pas toujours le cas.

IV.III.2. Analyse des blocs de synténie

Parmi les 27 paires d'orthologues potentiels retenus, 14 séquences forment 7 blocs de synténie répartis sur 6 LG de chicorée et 7 chromosomes de laitue (**figure 5.2** et **tableau 5.2**). Le **tableau 5.3** détaille la localisation et la taille de chacun des blocs mis en évidence ainsi que leur proportion par rapport au LG, au chromosome ainsi que par rapport à la taille totale du génome de la laitue et de la chicorée.

Tableau 5.2 : Correspondance entre les paires d'orthologues sur les LG de la chicorée et sur les chromosomes de la laitue. Gris : blocs de synténie observés

		Carte de la chicorée									
		LG1	LG2	LG3	LG4	LG5	LG6	LG7	LG8	LG9	
Carte de la laitue	Chromo.1	2					1		1		
	Chromo.2		2				1				
	Chromo.3			2							
	Chromo.4				2	1					
	Chromo.5		1			2					
	Chromo.6					1		1			
	Chromo.7									2	
	Chromo.8			1	1				1		
	Chromo.9				1	1	1			2	

Ces 7 blocs de synténie ont des tailles très variables allant de 0,3 cM à 68,8 cM chez la chicorée et de 0,9 à 118,7 Mpb chez la laitue. Un bloc de synténie peut recouvrir jusqu'à 48,1% d'un LG (bloc 7, LG9) et 60,7% d'un chromosome (Bloc 3, chromosome 3). Au total, les 7 zones de synténie recouvrent 145,8 cM (11,1%) du génome de la chicorée et 350,6 Mpb (13%) du génome de la laitue.

Vu que cette étude est basée sur la carte génétique de la chicorée et vu la taille des blocs, nous postulons qu'entre le génome de la chicorée et de la laitue, il existe des zones de macrosynténie. Cependant, la taille importante de certains blocs de synténie laisse supposer qu'il existe des réarrangements au sein de ceux-ci. Un nombre plus important d'orthologues dans ces régions permettrait de mettre en évidence ces réarrangements et donc d'affiner les zones de synténie entre les deux espèces.

Tableau 5.3 : Description des sept blocs de synténie observés sur la carte comparative entre la chicorée et la laitue à l'aide de 27 paires d'orthologues potentiels

	Bloc 1	Bloc 2	Bloc 3	Bloc 4	Bloc 5	Bloc 6	Bloc 7	Somme	Moyenne
Numéro de LG/ chromo.	1/1	2/2	3/3	4/4	5/5	9/9	9/7	-	-
Taille du bloc chez la chicorée (cM)	5,6	5,2	32,0	31,7	2,2	0,3	68,8	145,8	20,8
Taille du bloc chez la laitue (Mpb)	8,7	13,4	114,0	118,7	0,9	8,1	86,8	350,6	50,1
Proportion du LG (%)	4,9	3,5	26,7	21,9	2,6	0,2	48,1	-	15,4
Proportion du chromosome (%)	3,4	4,7	60,7	30,7	0,2	3,2	39,2	-	20,3
Proportion de la carte de chicorée (%)	0,4	0,4	2,4	2,4	0,2	0,02	5,2	11,1	1,6
Proportion de la carte de la laitue (%)	0,3	0,5	4,2	4,4	0,0	0,3	3,2	13,0	1,9

13 paires d'orthologues ne présentant pas de synténie entre les deux espèces, on peut émettre l'hypothèse que des réarrangements chromosomiques de type transversion, fusion et fission ont eu lieu au sein des deux génomes depuis leur divergence. Cela pourrait expliquer l'organisation de ces séquences. A ce jour, vu le peu d'éléments disponibles et vu que nous ne pouvons pas être certain que ce sont réellement des orthologues, cette hypothèse ne peut être approfondie. Afin d'avoir une meilleure idée de ces phénomènes, des marqueurs associés à des séquences doivent être ajoutés sur la carte comparative.

Les caractéristiques de chaque bloc sont détaillées :

- Bloc 1 :

Le bloc de synténie positionné sur le LG 1 et chromosome 1 possède une taille de 5,6 cM chez la chicorée et une taille estimée de 8,7 Mb chez la laitue. Cette région contient les deux marqueurs SNP (bip43 : EH707518 et bip21 : EH705277) délimitant le bloc. Cette partie du génome de la laitue a été étudiée dans différentes études de QTL que nous détaillerons plus bas.

- Bloc 2 :

Ce bloc est partagé entre le LG 2 et le chromosome 2. Cette zone représente 5,2 cM chez la chicorée et 13,4 Mb chez la laitue. Sur la carte génétique de la chicorée, cette zone possède les deux marqueurs SNP la délimitant (pa5 : FL675412 et bip37 : EH684149), un marqueur génomique de type SSR (GA357) et un marqueur AFLP.

- Bloc 3 :

Cette zone est située sur le LG 3 et le chromosome 3. Il recouvre 32 cM du groupe de liaison et 114 Mb du chromosome. C'est le bloc de synténie qui recouvre la plus grande proportion d'un chromosome soit 60,7%. Cette zone est constituée de huit marqueurs chez la chicorée : deux marqueurs SNP (bip41 : EL345653 et bip20 : EH691568), trois marqueurs microsatellites génomiques (GA036, GA118 et ATC002) et trois marqueurs AFLP (H1879, Ld et Kf).

▪ Bloc 4 :

Le bloc 4 est situé sur le LG4 de la chicorée et sur le chromosome 4 de la laitue. Il recouvre 31,7 cM du LG et 118,7 Mb du chromosome. C'est la zone de synténie la plus importante sur le génome de la laitue : ce bloc recouvre à lui seul 4,4% du génome total. Sur la carte génétique de la chicorée, on retrouve six marqueurs dans cette zone : un marqueur SNP (est40 : EH700985), deux SSR géniques (est8 : EH696752, est11 : FL675360), un SSR génomique (GA565) et deux marqueurs AFLP (A1192 et B1435).

▪ Bloc 5

Ce bloc est positionné sur le chromosome 5 de la laitue et sur le groupe de liaison 5 de la chicorée. Il recouvre 2,2 cM chez la chicorée et 0,9 Mb chez la laitue. C'est la plus petite zone de synténie sur la carte physique. Chez la chicorée, ce bloc est composé d'un SSR génique (ms1 : FL677058) et d'un marqueur SNP associé à la séquence FEH1(AJ242538), une fructane exohydrolase impliquée dans la dépolymérisation de l'inuline.

Chez la laitue, cette région est délimitée par une EST (CLSM14035) ayant deux orthologues potentiels chez la chicorée (Cifr07 : JQ082518 et FEH1 : AJ242538). Cette EST possède une très grande homologie avec FEH1 de la chicorée (identité de 243 acides aminés sur 286, e-Value de e^{-180}). L'homologie entre CLSM14035 et Cifr07 est moins évidente (e-Value e^{-80}). Suivant ces résultats de blast, nous considérerons que CLSM14035 est l'orthologue potentiel de FEH1 et non de Cifr07.

▪ Bloc 6

Ce bloc est commun au groupe de liaison 9 et au chromosome 9. Il possède une taille de 0,3 cM chez la chicorée et de 8,1 Mb chez la laitue. C'est le plus petit bloc de synténie sur la carte de la laitue. Il est composé uniquement des deux marqueurs SNP le délimitant (bip28 : EH709102 et bip57 : EH692696)

▪ Bloc 7 :

Ce dernier bloc est positionné sur le LG 9 de la chicorée et sur le chromosome 7 de la laitue. Il a une taille de 68,8 cM et de 86,9 Mb. C'est le plus grand bloc de synténie sur la carte de la chicorée

recouvrant 5,2% du génome. On y retrouve deux marqueurs SNP (bip33 : EH672649 et bip70 : FL679098), un SSR génomique (GA499) et cinq marqueurs AFLP (H1760, H1991, A2153, C1516 et H2211).

Les blocs de synténie 6 et 7 se trouvent sur le même groupe de liaison chez la chicorée. Par contre, ils ne se trouvent pas sur le même chromosome chez la laitue. Nous pouvons émettre l'hypothèse qu'un événement de translocation entre les chromosomes 9 et 7 s'est produit dans le génome de la laitue.

IV.III.3. Hypothèse sur le degré de synténie

Le degré de synténie entre deux espèces peut être très variable. En effet, plusieurs niveaux de synténie ont pu être observés au sein des différentes familles d'angiospermes. La famille des Brassicaceae dont fait partie l'espèce modèle *A. thaliana* montre un grand nombre de réarrangements entre les différents genres. Le degré de synténie semble donc faible pour cette famille, principalement entre le genre *Brassica* et *A. thaliana* (Schranz et al. 2006). Les espèces appartenant à la famille des Poaceae (graminées) montrent une grande diversité au niveau de la taille des génomes et du caryotype. Malgré cela, il semblerait qu'il y ait une très grande conservation de l'organisation des génomes dans cette famille qui semble donc, en opposition aux Brassicaceae, hautement synténique (Bolot et al. 2009b). L'analyse de la synténie au sein des Fabaceae (légumineuses) montre des résultats intermédiaires. En effet, il existe une grande conservation de l'organisation des génomes d'espèces proches et le degré de synténie semble diminuer avec la distance phylogénétique (Choi et al. 2004).

Notre cartographie comparée entre la chicorée industrielle et la laitue réalisée à l'aide de 27 paires d'orthologues potentiels a permis de mettre en évidence sept blocs de synténie à l'aide de 14 séquences et probablement un événement de translocation. Ces blocs sont localisés sur six des neuf groupes de liaison de la chicorée et sur sept des neuf chromosomes de la laitue et recouvrent environ 12 % des génomes. On peut émettre l'hypothèse que les 13 marqueurs qui ne sont pas localisés dans ces blocs de synténie ont subi des événements de translocation par rapport à leurs chromosomes d'origine. Cependant, le fait que ces « événements » soient constitués d'un seul marqueur ne permet pas d'appuyer cette supposition.

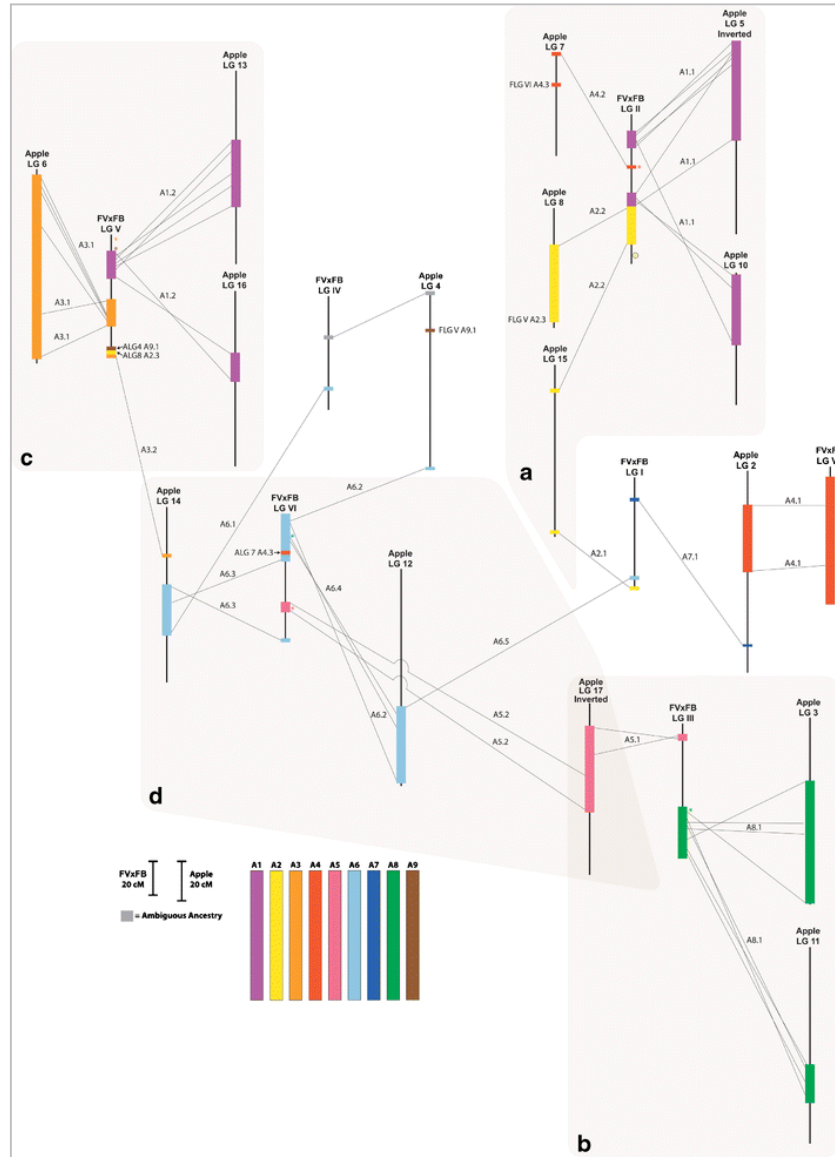


Figure 5.3 : Représentation des régions synténiques entre *Malus* (2n=34) et *Fragaria* (2n=14) déterminées par 56 marqueurs orthologues. Les couleurs correspondent aux différents chromosomes ancestraux. Le génome ancestral est composé de neuf chromosomes allant de A1 à A9. FVxFB LG I à LG VII : groupes de liaison de la carte génétique issue du croisement entre *F. vesca* et *F. bucharica*. Apple LG 1 à 17 : Groupes de liaison de la carte génétique du genre *Malus* (Bushakra et al. 2011)

Il est possible d'évaluer le niveau de synténie en calculant le taux de réarrangements entre les espèces étudiées. Le taux de réarrangements = (le nombre de réarrangements chromosomiques observés) / (2x le temps de divergence entre les deux espèces). Le temps de divergence est multiplié par deux car les événements observés peuvent avoir eu lieu lors de l'évolution des deux génomes (Burke et al. 2004). Concernant notre cartographie comparée, le nombre de réarrangements n'a pas pu être évalué. De plus, le temps de divergence entre la laitue et la chicorée n'est pas connu. Le taux de divergence ne peut donc pas être calculé. Puisque nous ne pouvons quantifier le niveau de synténie, nous allons comparer les résultats obtenus avec une autre étude de cartographie comparée présentant des blocs de synténie délimités par plus de deux marqueurs mais également délimités par un seul, voire deux marqueurs comme dans notre cas.

Une cartographie comparée a été réalisée au sein de la famille des Rosaceae entre les genres *Malus* (la pomme) et *Fragaria* (*F. vesca*, la fraise des bois) (Bushakra et al. 2011). Dans cette étude, 56 marqueurs orthologues ont pu mettre en évidence 21 blocs de synténie entre les deux espèces (**figure 5.3**). Cela représente 49 % des 588 cM du génome de la pomme et 71 % des 1.277 cM du génome de la fraise. Parmi les 21 blocs de synténie, six sont composés de plus de deux marqueurs, quatre possèdent deux marqueurs et 11 blocs sont constitués d'un seul marqueur. Le nombre de régions synténiques possédant au moins deux marqueurs (10 au total) est semblable à ce qui est observé sur notre carte. L'allure de la carte de Bushakra est similaire à la nôtre dans le sens où l'organisation des groupes de liaison est fortement divergente entre les deux espèces. Au final, un nombre deux fois plus élevé de marqueurs a permis d'obtenir six blocs de synténie possédant plus de deux marqueurs orthologues, ce qui permet de mieux appuyer la validité des observations. Sur notre carte comparative, des réarrangements chromosomiques potentiels sont mis en évidence par un seul marqueur. Le nombre d'orthologues entre les deux cartes doit donc être augmenté afin de vérifier ces hypothèses et de confirmer les différents blocs de synténie. Notre carte comparative, la première du genre au sein de la tribu des Lactuceae (Cichorieae), est donc un point d'ancrage pour une future étude de synténie à plus grande échelle.

IV.III.4. Transfert de données entre la chicorée et la laitue

En dehors de l'étude de l'évolution des génomes, l'intérêt de la cartographie comparée réside dans le fait qu'il est possible de transférer des données génétiques entre régions synténiques d'une espèce à l'autre.

Le développement et l'accessibilité des cartes génétiques et physiques de la laitue au sein du Compositeae Genome Project permettent, d'une étude à l'autre, de se référer à la même source de données et donc de connaître facilement la localisation des séquences utilisées dans les différentes publications. A ce jour, plusieurs études se réfèrent à ces cartes de laitue (Hayes et al. 2011; McHale et al. 2009; Schwember and Bradford 2010; Simko et al. 2011; Timms et al. 2006). Trois d'entre elles présentent des QTL ou des loci phénotypiques au sein des blocs de synténie mis en évidence dans notre cartographie comparée :

- Co-localisation de QTL/loci phénotypiques avec le bloc 1 de synténie

Le bloc 1 de synténie est délimité chez la chicorée par deux marqueurs co-dominants, bip43 : EH707518 et bip21 : EH705277. Ces marqueurs sont associés à une séquences EST dont la fonction n'est pas connue pour bip43 et dont la fonction est associée à une protéine liant la Rubisco pour bip21 (Rubisco subunit binding-protein alpha subunit). Ce bloc de synténie est co-localisé avec six QTL/ loci phénotypiques provenant de trois publications différentes.

Tout d'abord, une carte génétique de laitue a été réalisée entièrement avec des gènes-candidats associés à la résistance aux pathogènes (McHale et al. 2009). Plusieurs dizaines de loci phénotypiques de résistance ainsi que des QTL ont été positionnés sur cette carte. Le bloc 1 de synténie de notre cartographie comparée est co-localisé avec deux loci phénotypiques LDm43 et LDm17. Ces loci dominants entraînent, suivant leur génotype, une résistance complète de la laitue à *Bremia lactucae*, l'oomycète responsable du mildiou, une maladie foliaire présente chez la laitue. Les mécanismes ainsi que les séquences génétiques associés à cette résistance ne sont pas connus.

Ensuite, les deux EST du bloc 1 de la laitue ont été utilisées lors d'une étude de synténie entre *A. thaliana* et *L. sativa* (ortho.Bip21=LE1066 et

ortho.Bip43=LK1303) (**figure 5.4**) (Timms et al. 2006). Dans cette étude, certains gènes d'intérêt ont été sélectionnés chez *A. thaliana* et les séquences présentes dans une région de 0,5 Mb autour du gène ont été utilisées afin de trouver des orthologues chez la laitue. La région qui nous intéresse est associée chez *A. thaliana* au gène LEAFY. Ce gène code pour un facteur de transcription contrôlant l'identité du méristème. Cette protéine déclenche le développement des cellules non différenciées du méristème pour la formation des fleurs (Weigel et al. 1992). Ce facteur est donc impliqué dans la floraison.

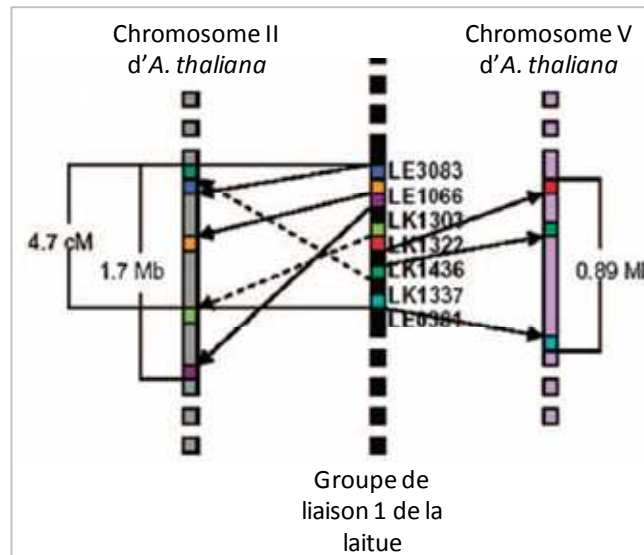


Figure 5.4 : Représentation d'une partie de la région LEAFY chez *A. thaliana* (chromosome 2 et 5) ainsi que chez la laitue (LG1). Les marqueurs LE1066 et LK1303 sont les orthologues potentiels des marqueurs bip21 et bip43 présents sur le LG1 de la carte génétique de la chicorée. Les flèches pleines indiquent l'emplacement du meilleur résultat et les flèches en pointillés indiquent le second résultat obtenu lors du tBlastx (Timms et al. 2006)

L'analyse de la syntenie dans cette région montre une certaine conservation entre la laitue et *A. thaliana*. Cette zone recouvre environ 1,7 Mb du chromosome deux chez *A. thaliana* et se compose de cinq séquences (LE3083, LE1066, LK1303, LK1322 et LE0381). L'ordre de trois de ces séquences dont font partie nos deux orthologues est conservé entre les deux espèces. Il semblerait cependant que cette région soit dupliquée chez *A. thaliana* puisque plusieurs séquences ont également été retrouvées sur le chromosome 5 (**figure 5.4**). Le bloc 1 de syntenie mis en évidence lors de notre cartographie

comparée fait donc partie de la région LEAFY mis en évidence dans cet article. Nous pouvons émettre l'hypothèse que cette région du génome de la chicorée est potentiellement liée au gène LEAFY et donc associé à la floraison.

Enfin, trois QTL (Slg1.1, SlgP1.1, SlgPCD1.1) associés à la longévité de la graine suivant les conditions de stockage ont été co-localisés avec le bloc 1 de synténie chez la laitue (Schwember and Bradford 2010).

- Co-localisation d'un locus phénotypique sur le bloc 3 de synténie

Le bloc trois de notre cartographie comparée a été localisé sur la carte de la laitue dont les séquences sont associées à la résistance aux pathogènes dont nous avons parlé ci-dessus (McHale et al. 2009). L'orthologue potentiel du marqueur bip41(EL345653) se trouve dans la même région qu'un locus de résistance *cor* sur le chromosome 3 de la laitue. Ce locus phénotypique confère, de manière récessive, une résistance à *Rhizomonas suberifaciens*, une bactérie responsable de la maladie dite de la racine liégeuse (corky root) (Moreno-Vazquez et al. 2003). À ce jour, ce locus n'a pas encore été caractérisé. Cinq EST sont localisées au sein du locus *cor* sur la carte de MCHale. Celles-ci ont été blastées contre les EST de chicorée donnant cinq Best BLAST Hit (**tableau 5.4**).

Tableau 5.4 : Résultat du tBlastX entre les EST localisées au sein ou près du locus récessif *Cor* chez la laitue et les EST de chicorée

Séquences localisées dans ou près du locus <i>Cor</i>	Source	N°d'accession du Best BLAST Hit contre les EST de chicorée	e-Value
CLX_S3_contig11134	McHale et al. 2009	FL673046.1	8,00E-63
CLX_S3_contig1242	McHale et al. 2010	EH674084.1	1,00E-59
CLX_S3_contig3740	McHale et al. 2011	FL673046.1	8,00E-42
CLX_S3_contig535	McHale et al. 2012	EH697790.1	3,00E-113
CLX_S3_contig9203	McHale et al. 2013	EH692065.1	2,00E-58
AY207411.1	Moreno-Vazquez et al. 2002	EH695118	2,00E-59

En 2002, six marqueurs co-dominants localisés dans la même région que le locus *cor* ont été développés afin d'identifier et de sélectionner les individus résistants à cette maladie (Moreno-Vazquez et al. 2003). Les séquences associées à ces marqueurs sont disponibles. Suite à

l'analyse BLAST de celles-ci avec les EST de chicorée, une de ces six EST (AY207411) a permis d'identifier une séquence similaire chez la chicorée (EH695118) avec une e-value de 2,00E-59.

Au final, les six EST de chicorée obtenues par BLAST (**tableau 5.4**) sont des candidats potentiellement intéressants pour l'étude de la résistance aux pathogènes chez la chicorée.

De plus, le marqueur bip41 co-localisé avec le locus *cor* se trouve au sein d'une EST dont la fonction est associée à une coniferylalcohol 5-hydroxylase, une enzyme impliquée dans la synthèse de la lignine. De par sa fonction potentielle et sa localisation, la séquence associée au marqueur bip41 pourrait également jouer un rôle dans la résistance aux pathogènes.

- Co-localisation d'un QTL avec le bloc 4 de synténie

Le bloc 4 de synténie comprenant quatre marqueurs co-dominants chez la chicorée est localisé sur la carte physique de la laitue dans la même région que le QTL Slg4.2 (Schwember and Bradford 2010) associé à la longévité des graines en fonction des conditions de stockage.

Tableau 5.5 : Récapitulatif des QTL et des loci phénotypiques localisés chez la laitue au sein des blocs de synténie ainsi que des marqueurs codominants de chicorée localisés dans ou à proximité des QTL/ loci

Bloc de synténie	Nom du QTL/ locus chez la laitue	Description	Source	Marqueurs codominants de la chicorée localisés à proximité des QTL/loci
1	LDm43	Locus dominant de résistance au Mildiou (<i>Bremia lactucae</i>)	McHale et al. 2009	bip43: predicted protein bip21: rubisco subunit binding-protein alpha subunit
	LDm17	Locus dominant de résistance au Mildiou (<i>Bremia lactucae</i>)	McHale et al. 2009	
	Région LEAFY	LEAFY: facteur de transcription contrôlant l'identité du méristème	Timms et al. 2006	
	Slg1.1	QTL associé à la longévité de la graine	Schwember et al. 2010	
	SlgP1.1	QTL associé à la longévité de la graine	Schwember et al. 2010	
	SlgPCD1.1	QTL associé à la longévité de la graine	Schwember et al. 2010	
3	cor	Locus récessif de résistance à <i>Rhizomonas suberifaciens</i> responsable de la racine liègeuse	McHale et al. 2009	bip41: coniferylalcohol 5-hydroxylase GA036: SSR génomique GA118: SSR génomique
4	Slg4.2	QTL associé à la longévité de la graine	Schwember et al. 2010	est8: trehalose-6-phosphate synthase est11: glycosyltransferase, CAZy family GT8 est40: lactoylglutathione lyase GA565: SSR génomique

Le **tableau 5.5** résume les différents QTL/loci phénotypiques localisés au sein de nos blocs de synténie ainsi que les marqueurs de chicorée présents dans ces zones. Vu la proximité physique de neuf marqueurs co-dominants de chicorée avec huit QTL/loci phénotypiques, nous pouvons émettre l'hypothèse que ces marqueurs pourraient être liés soit directement, soit indirectement à des phénotypes d'intérêt chez la chicorée. Afin de vérifier cette supposition, ces marqueurs devront être utilisés dans le cadre d'étude d'association.

Le bloc 1 de synténie possède chez la laitue trois QTL, deux loci phénotypiques et une région associée au gène LEAFY. Cette zone semble donc posséder des gènes impliqués dans plusieurs mécanismes tels que la floraison ou encore la résistance au pathogène. Ce bloc de synténie couvre 5.6 cM du génome de la chicorée, ce qui représente suivant nos estimations (1 cM= 1 Mb) environ 5.6 Mb. Le séquençage imminent du génome de *C. intybus* nous permettra de trouver, à l'aide des marqueurs bordant le bloc1, les séquences localisées au sein de cette zone. Cela permettra de caractériser les gènes impliqués dans les mécanismes associés aux QTL et aux loci phénotypiques.

La cartographie comparée a donc permis, grâce à la mise en évidence de régions synténiques, de mettre en lumière des régions du génome ainsi que des marqueurs potentiellement impliqués dans divers mécanismes d'intérêt pour la sélection de la chicorée.

V. CONCLUSION

La chicorée est une espèce dont les ressources génétiques telles que les QTL, les banques de séquences, les loci phénotypiques, sont relativement faibles. En revanche, les données génétiques de la laitue se développent de manière importante dans le cadre du Compositea Genome Project. Une carte physique constituée de plus de 50.000 contigs d'EST ainsi que de nombreuses études de QTL et de loci phénotypiques sont actuellement disponibles. Afin de pouvoir transférer les données génétiques d'une espèce à l'autre, nous avons élaboré une cartographie comparée entre la chicorée et la laitue.

Dans un premier temps, la recherche de paires d'orthologues par la méthode d'alignement tBlastx et par la recherche de RBH entre les 39

séquences positionnées sur la carte de la chicorée avec une partie des EST de la carte physique et génétique de la laitue a permis d'obtenir 27 paires d'orthologues potentiels. Celles-ci ont été utilisées afin de construire la carte comparative.

Cette cartographie comparée a permis de mettre en évidence sept blocs de synténie de taille moyenne de 20,8 cM chez la chicorée et de 50 Mb chez la laitue macrosynténie. Cela représente 11 % du génome de la chicorée et 13 % du génome de la laitue. Ces blocs sont délimités à l'aide de 14 paires d'orthologues potentiels. Un éventuel événement de translocation a également pu être mis en évidence. La localisation éparse des 13 marqueurs restant pourrait être due à des réarrangements chromosomiques. Cependant, cette hypothèse devra être vérifiée avec un nombre de marqueurs plus élevé.

Ces blocs de synténie ont permis de faire un lien supposé entre des QTL et des loci phénotypiques de la laitue avec des zones du génome de la chicorée. Au total, quatre QTL associés à la longévité des graines, trois loci phénotypiques associés à la résistance aux pathogènes et une région associée à un gène impliqué dans la floraison ont été localisés dans des régions synténiques.

Au final, cette étude a permis de mettre en évidence de la macrosynténie entre la laitue et la chicorée pour un peu plus du dixième de leur génome. De plus, cette analyse a permis d'identifier des régions du génome de la chicorée ainsi que des marqueurs pouvant jouer un rôle important dans les variations phénotypiques de la chicorée. En conclusion, cette cartographie comparée doit être perçue comme un point d'ancrage pour une plus vaste étude de synténie entre les deux espèces. Cette première approche nous a permis de voir le potentiel de ce genre d'étude.

Conclusion et perspectives

La sélection de la chicorée industrielle qui, de façon générale, se fait de manière phénotypique, peut être facilitée par l'utilisation d'outils moléculaires tels que les marqueurs ou encore la cartographie.

Afin de développer ces outils, nous avons contribué lors de cette thèse, en collaboration avec l'ECAV de l'UCL et la société Chicoline, à la construction d'une carte génétique de la chicorée à l'aide de marqueurs dominants de type AFLP et co-dominants de types SSR et SNP. L'utilisation de marqueurs associés à des séquences transcrites nous a permis d'initier une cartographie comparée avec la laitue, une espèce proche dont les ressources génétiques sont en pleine expansion.

La première étape de ce travail consistait à rechercher du polymorphisme au sein de séquences transcrites afin de développer des marqueurs moléculaires de types SSR et SNP. Pour ce faire, une analyse bioinformatique des banques EST disponibles pour la chicorée industrielle a été réalisée.

La recherche de motifs SSR a mis en évidence que 15 % des EST analysées comportaient ces motifs dont 59 % étaient de type trinuécléotidique. Suite à la sélection de ces séquences suivant leur fonction liée à des caractères d'intérêt chez la chicorée et suivant les caractéristiques intrinsèques des SSR, 109 séquences ont été retenues.

La mise en évidence de SNP au sein de la banque EST Phytomol a permis de retenir 64 séquences contenant un total de 294 SNP localisés dans les régions 3'-UTR et les régions codantes. Comme attendu, ces mutations étaient pour plus de 60 % des cas, des transitions et, au sein des régions codantes, 85 % sont des mutations synonymes.

Au total, 173 séquences ont ainsi été retenues pour le développement de marqueurs de types SSR et SNP.

La deuxième étape consistait à rechercher au sein de séquences d'intérêt, du polymorphisme présent dans la population de ségrégation SAMCHIC réalisée par la société Cosucra et utilisée pour la construction de la carte génétique. Ces séquences d'intérêt étaient divisées en plusieurs catégories : des séquences potentiellement orthologues de la laitue, une partie des séquences cartographiées sur une autre carte de chicorée (Cadalen et al.2010) et des séquences

associées à des fonctions d'intérêt dans des processus tels que la pathogénèse et la floraison. Au total, 315 séquences ont été analysées et 94 ont été retenues car elles présentaient du polymorphisme SNP dans la population SAMCHIC.

Au final, l'analyse des banques EST et de séquences d'intérêt a permis de mettre en évidence 693 sites de types SSR et SNP au sein de 267 séquences.

La troisième étape de ce travail consistait à développer des marqueurs moléculaires au sein de ces séquences sélectionnées. Concernant les SSR, suite aux mises au point, 70 marqueurs valides ont été développés. À propos des marqueurs SNP, l'utilisation de méthodes de génotypage de type PAMSA et Bi-PASA a permis de concevoir 60 marqueurs valides. Au total, un jeu de 130 marqueurs co-dominants et associés à des séquences transcrites est donc à ce jour disponible chez la chicorée industrielle.

Afin de réaliser la quatrième étape, c'est-à-dire la construction de la carte génétique, ces 130 marqueurs ont été utilisés pour génotyper la population SAMCHIC. Au final, 14 marqueurs SSR et 27 marqueurs SNP soit 41 marqueurs ont été retenus car ils présentaient du polymorphisme dans le croisement.

Dans le but de saturer la carte, des marqueurs AFLP ont été utilisés. Cette méthode a amené de nombreux problèmes liés au génotypage. Seuls 13 marqueurs AFLP ont pu être retenus pour la cartographie.

Au total, 54 marqueurs ont été utilisés pour le génotypage de 200 individus de la population de ségrégation SAMCHIC. Ces marqueurs complètent le jeu de 163 marqueurs AFLP développés par l'ECAV à l'UCL ainsi que 26 marqueurs SSR génomiques provenant de la société Chicoline.

Au final, la carte génétique de la chicorée s'est basée sur 237 marqueurs (152 marqueurs AFLP, 27 marqueurs SNP, 28 marqueurs SSR génomiques et 12 marqueurs EST-SSR). Neuf groupes de liaison ont ainsi été formés avec une taille totale de 1208 cM, ce qui représente 92 % de la taille totale estimée du génome. Les marqueurs sont à une distance moyenne de 5,6 cM. Si l'on considère que le génome de la chicorée comprend 1.333 Mb, cela signifie que sur notre carte, 1cM représente environ 1Mb.

67 marqueurs co-dominants ont été positionnés sur cette carte dont 39 sont associés à des séquences exprimées et 12 sont des orthologues potentiels de séquences cartographiées sur la carte génétique de la laitue.

La cinquième et dernière étape de cette thèse consistait à réaliser la cartographie comparée entre la carte génétique de la chicorée et la carte physique de la laitue. Pour ce faire, nous avons repris toutes les séquences exprimées positionnées sur la carte de la chicorée et nous avons entrepris de chercher les orthologues potentiels positionnés sur la carte physique de la laitue. 27 paires d'orthologues ont ainsi été retenues et utilisées pour la comparaison des deux génomes. Sept blocs de synténie en moyenne de 20,8 cM chez la chicorée et de 50 Mb chez la laitue ont pu être mis en évidence. Cette macrosynténie recouvre 11 % du génome de la chicorée et 13 % du génome de la laitue.

Quatre QTL, trois loci phénotypiques et une région « LEAFY » ont été localisés dans ces zones synténiques chez la laitue. Les marqueurs de chicorée co-localisés avec ces loci peuvent être considérés comme des gènes candidats positionnels pouvant être utilisés dans de futures études d'association.

Ce travail a donc permis au final de développer des outils pour la sélection assistée par marqueurs de la chicorée industrielle. Les outils ainsi mis en place peuvent être intégrés dans les étapes de sélection et dans de futurs projets de recherche. Les différentes perspectives sont détaillées ci-dessous.

- Utilisation des 130 marqueurs SSR et SNP chez les variétés cultivées.

Afin d'analyser les différents cultivars de chicorée industrielle, les 70 marqueurs moléculaires de type SSR peuvent être utilisés à différentes fins pour la sélection telles que :

- l'évaluation de la diversité génétique afin d'identifier les populations possédant des ressources génétiques importantes,
- l'identification des individus via les allèles ou les combinaisons d'allèles spécifiques des différents cultivars,

- l'analyse des génétiques pour identifier des populations génétiquement distantes dans le but de créer de nouveaux croisements.

De plus, les marqueurs SNP et SSR peuvent également être utilisés afin de géotyper une nouvelle population de ségrégation dans le but de développer des QTL.

- Utilisation de la carte génétique pour initier l'assemblage du génome

Le séquençage complet du génome d'un individu homozygote de chicorée est en cours dans le cadre du projet INULONG (projet DGARNE : Subvention D31-1221) en partenariat avec l'ECAV de l'UCL, la société Chicoline (division de Cosucra groupe Warcoing S.A.) et l'URBV des FUNDP. Ce séquençage à haut débit va permettre d'obtenir des contigs non-ordonnés de plusieurs centaines de kb recouvrant entre 95 et 99 % du génome. La recherche au sein des contigs des séquences positionnées sur la carte génétique va permettre de localiser une partie du génome sur les groupes de liaison. La carte génétique pourra être utilisée afin d'initier la construction de la carte physique de la chicorée industrielle. Pour l'instant, une quarantaine de marqueurs peut être identifiée au sein du génome. Une densité de marqueurs plus importante sur la carte de la chicorée permettrait d'augmenter les points d'ancrage des contigs sur la carte.

- Approfondir l'étude de synténie entre la chicorée et la laitue

La localisation d'une partie du génome sur la carte génétique permettrait d'approfondir l'étude de synténie entre la chicorée et la laitue. Les contigs ainsi localisés pourraient être utilisés pour une analyse RBH avec les contigs positionnés sur la carte physique de la laitue. Cette nouvelle source d'information permettrait d'approfondir l'étude de l'évolution des génomes au niveau de la macrosynténie mais également au niveau de la microsynténie.

- Utilisation des marqueurs pour les études d'associations

De manière générale, les marqueurs co-dominants développés lors de cette thèse peuvent être utilisés au sein d'études d'associations

puisque le polymorphisme mis en évidence est potentiellement lié directement ou indirectement à des variations phénotypiques.

De plus, puisque chaque marqueur est associé à une séquence transcrite, une sélection préalable des marqueurs analysés peut être réalisée suivant leur fonction par rapport aux critères phénotypiques étudiés dans l'étude d'association.

Enfin, neuf marqueurs de chicorée ont été identifiés au sein de régions synténiques possédant des QTL et des loci phénotypiques chez la laitue. Ce sont donc des marqueurs de choix pour de futures études d'association liées aux problématiques associées à ces QTL et loci phénotypiques.

- Caractérisation du bloc 1 de synténie

Le bloc 1 de synténie mis en évidence entre la laitue et la chicorée est délimité par les marqueurs bip43 (EH707518) et bip21 (EH705277). Cette région semble être, chez la laitue, associée à plusieurs mécanismes tels que la résistance aux pathogènes, à la longévité de la graine ou à la floraison. Il serait donc intéressant d'envisager une étude approfondie des séquences et du polymorphisme présents dans cette zone du génome chez la chicorée.

Les objectifs définis pour cette thèse ont donc été atteints : une carte génétique a été construite, des marqueurs associés à des séquences transcrites ont été développés et une cartographie comparée avec la laitue a été réalisée. Grâce à ces travaux, de nombreuses perspectives peuvent ainsi être envisagées pour l'étude et la sélection de la chicorée industrielle.

Bibliographie

Abajian C (1994) Sputnik.

Abravaya K, Carrino JJ, Muldoon S, Lee HH (1995) Detection of point mutations with a modified ligase chain reaction (Gap-LCR). *Nucleic Acids Research* 23:675-682

AGI (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796-815

Altman A, Hasegawa PM, ScienceDirect (2011) Plant biotechnology and agriculture prospects for the 21st century. Academic, Oxford

Argyris J, Truco M, Ochoa O, McHale L, Dahal P, Van Deynze A, Michelmore R, Bradford K (2011) A gene encoding an abscisic acid biosynthetic enzyme (*LsNCED4*) collocates with the high temperature germination locus *Htg6.1* in lettuce (*Lactuca* sp.). *TAG Theoretical and Applied Genetics* 122:95-108

Arzate A (2005) Extraction du sucre de betterave. Centre de recherche, de développement et de transfert technologique en acériculture (Le Centre **ACER** Inc.)

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* 3:e3376

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH (2008) Multiple Paleopolyploidizations during the Evolution of the Compositae Reveal Parallel Patterns of Duplicate Gene Retention after Millions of Years. *Molecular Biology and Evolution* 25:2445-2455

Beckmann JS, Soller M (1983) Restriction fragment length polymorphisms in genetic improvement: methodologies, mapping and costs. *TAG Theoretical and Applied Genetics* 67:35-43

Bennett MD, Bhandol P, Leitch IJ (2000) Nuclear DNA Amounts in Angiosperms and their Modern Uses-807 New Estimates. *Annals of Botany* 86:859-909

Bolot S, Abrouk M, Masood-Ouraishi U, Stein N, Messing J, Feuillet C, Salse J (2009) The "inner circle" of the cereal genomes. *Current Opinion in Plant Biology* 12:119-125

- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314-331
- Brunet FdrG, Crollius HR, Paris M, Aury J-M, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M (2006) Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes. *Molecular Biology and Evolution* 23:1808-1816
- Burke JM, Lai Z, Salmaso M, Nakazato T, Tang S, Heesacker A, Knapp SJ, Rieseberg LH (2004) Comparative Mapping and Rapid Karyotypic Evolution in the Genus *Helianthus*. *Genetics* 167:449-457
- Burrell AM, Taylor KG, Williams RJ, Cantrell RT, Menz MA, Pepper AE (2011) A comparative genomic map for *Caulanthus amplexicaulis* and related species (Brassicaceae). *Molecular Ecology* 20:784-798
- Bushakra J, Sargent D, Cabrera A, Crowhurst R, Girona E, Velasco R, Symonds V, van der Knaap E, Troggio M, Gardiner S, Chagné D (2011) Rosaceae conserved orthologous set (RosCOS) markers as a tool to assess genome synteny between *Malus* and *Fragaria*. *Tree Genetics & Genomes*:1-16
- Caballero A, Quesada H (2010) Homoplasy and Distribution of AFLP Fragments: An Analysis In Silico of the Genome of Different Species. *Molecular Biology and Evolution* 27:1139-1151
- Cadalen T, Mörchen 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
- Cai H-Y, Diao S, He Y-G, Zhang L-P, Liu S-J, Zhu Y-G, Zhang Z-H (2012) Genetic and physical mapping of qHY-8, a pleiotropic QTL for heading date and yield-related traits in rice. *Euphytica* 184:109-118
- Cao X, Zhou J, Gong X, Zhao G, Jia J, Qi X (2012) Identification and Validation of a Major Quantitative Trait Locus for Slow-rusting Resistance to Stripe Rust in Wheat. *Journal of Integrative Plant Biology*:no-no
- CGP (2012) Compositea Genome Project Database <http://compgenomics.ucdavis.edu/>.

Chakravarti A, Lasher LK, Reefer JE (1991) A maximum likelihood method for estimating genome length using genetic linkage data. *Genetics* 128:175-182

Chapman M, Hvala J, Strever J, Matvienko M, Kozik A, Michelmore R, Tang S, Knapp S, Burke J (2009) Development, polymorphism, and cross-taxon utility of EST-SSR markers from safflower (*Carthamus tinctorius* L.). *TAG Theoretical and Applied Genetics* 120:85-91

Cheema J, Dicks J (2009) Computational approaches and software tools for genetic linkage map estimation in plants. *Briefings in Bioinformatics* 10:595-608

Cheema J, Ellis THN, Dicks J (2010) THREaD Mapper Studio: a novel, visual web server for the estimation of genetic linkage maps. *Nucleic Acids Research* 38:W188-W193

Choi H-K, Mun J-H, Kim D-J, Zhu H, Baek J-M, Mudge J, Roe B, Ellis N, Doyle J, Kiss GB, Young ND, Cook DR (2004) Estimating genome conservation between crop and model legume species. *Proceedings of the National Academy of Sciences of the United States of America* 101:15289-15294

Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363:557-572

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. *The Plant Journal* 37:778-786

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674-3676

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* 36:e122

Dauchot N (2007) EST derived molecular markers applied to polyfructan metabolism in *Chichorium intybus*. Dissertation for the requirement for the degree of PhD in biological Sciences. University of Namur

Dauchot N, Mingot D, Purnelle B, Muys C, Watillon B, Boutry M, Van Cutsem P (2009) Construction of 12 EST libraries and characterization of a 12,226 EST dataset for chicory (*Cichorium intybus*) root, leaves and nodules in the context of carbohydrate metabolism investigation. *BMC Plant Biol* 9:14

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

de Vienne D (1998) Les marqueurs moléculaires en génétique et biotechnologies végétales. INRA, Paris

Dear PH (2007) Bioinformatics, Scion Publishing Limited edn

Decreux A, Messiaen J (2005) Wall-associated Kinase WAK1 Interacts with Cell Wall Pectins in a Calcium-induced Conformation. *Plant and Cell Physiology* 46:268-278

Dielen V, Notté C, Lutts S, Debavelaere V, Van Herck J-C, Kinet J-M (2005) Bolting control by low temperatures in root chicory (*Cichorium intybus* var. *sativum*). *Field Crops Research* 94:76-85

Dognaux S (2009) Sélection assistée par marqueurs chez *Cichorium intybus* L.: recherche de marqueurs moléculaires ciblant les gènes de floraisons. Mémoire présenté pour l'obtention du grade académique de master en biochimie et biologie moléculaire et cellulaire. Facultés Universitaires Notre-Dame de la Paix, Namur

Doré C, Varoquaux F (2006) Histoire et amélioration de cinquante plantes cultivées. Editions Quae

Dresselhaus T, Cordts S, Heuer S, Sauter M, Lörz H, Kranz E (1999) Novel ribosomal genes from maize are differentially expressed in the zygotic and somatic cell cycles. *Molecular and General Genetics MGG* 261:416-427

Dussle CD, Quint MQ, Xu MX, Melchinger AM, Lübberstedt TL (2002) Conversion of AFLP fragments tightly linked to SCMV resistance genes into simple PCR-based markers. *TAG Theoretical and Applied Genetics* 105:1190-1195

Eckardt NA (2001) Everything in Its Place: Conservation of Gene Order among Distantly Related Plant Species. *The Plant Cell Online* 13:723-725

- Edelman J, Jefford TG (1968) The mechanism of fructosan metabolism in higher plants as exemplified in *Helianthus tuberosus*. *New Phytologist* 67:517-531
- Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in plant biology. *American Journal of Botany* 99:175-185
- Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* 16:551-558
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* 6:e19379
- Emebiri LC (2010) An EST-SSR Marker Tightly Linked to the Barley Male Sterility Gene (*msg6*) Located on Chromosome 6H. *Journal of Heredity* 101:769-774
- Fan J-B, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelplnagel J, Chee MS (2003) Highly Parallel SNP Genotyping. *Cold Spring Harbor Symposia on Quantitative Biology* 68:69-78
- Fishman L, Kelly AJ, Morgan E, Willis JH (2001) A genetic map in the *Mimulus guttatus* species complex reveals transmission ratio distortion due to heterospecific interactions. *Genetics* 159:1701-1716
- Fitch WM (2000) Homology a personal view on some of the problems. *Trends Genet* 16:227-231
- Franck A (2002) Technological functionality of inulin and oligofructose. *The British journal of nutrition* 87 Suppl 2:S287-291
- Fujimori S, Washio T, Higo K, Ohtomo Y, Murakami K, Matsubara K, Kawai J, Carninci P, Hayashizaki Y, Kikuchi S, Tomita M (2003) A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. *FEBS Letters* 554:17-22
- Funk V, Bayer R, Keeley S, Chan R, Watson L, Gemeinholzer B, Schilling E, Panero J, Baldwin B, Garcia-Jacas N, Susanna A, Jansen R (2005) Everywhere but Antarctica: Using a supertree to understand the diversity and distribution of the Compositae. *Biol Skr* 55:343-374

- Garrett R, Grisham C (2000) Biochimie, de Boeck Université edn
- Gaudet M, Fara A-G, Sabatti M, Kuzminsky E, Mugnozza G (2007) Single-reaction for SNP Genotyping on Agarose Gel by Allele-specific PCR in Black Poplar (*Populus nigra* L.). Plant Molecular Biology Reporter 25:1-9
- Gibson G, Muse S (2004) Précis de génomique, de Boeck edn
- Gibson GR, McCartney AL, Rastall RA (2005) Prebiotics and resistance to gastrointestinal infections. Br J Nutr 93 Suppl 1:S31-34
- Gibson GR, Roberfroid MB (1995) Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. The Journal of nutrition 125:1401-1412
- Grantham R (1974) Amino Acid Difference Formula to Help Explain Protein Evolution. Science 185:862-864
- Graur D, Li W-H (2000) Fundamentals of molecular Evolution. Sinauer Associates
- Guyon-Debast A, Lecureuil A, Bonhomme S, Guerche P, Gallois J-L (2010) A SNP associated with alternative splicing of RPT5b causes unequal redundancy between RPT5a and RPT5b among Arabidopsis thaliana natural variation. BMC Plant Biology 10:158
- Hayashi E, You Y, Lewis R, Calderon M, Wan G, Still D (2012) Mapping QTL, epistasis and genotype × environment interaction of antioxidant activity, chlorophyll content and head formation in domesticated lettuce (*Lactuca sativa*). TAG Theoretical and Applied Genetics 124:1487-1502
- Hayes R, McHale L, Vallad G, Truco M, Michelmore R, Klosterman S, Maruthachalam K, Subbarao K (2011) The inheritance of resistance to Verticillium wilt caused by race 1 isolates of "*Verticillium dahliae*" in the lettuce cultivar La Brillante. TAG Theoretical and Applied Genetics 123:509-517
- Heesacker A, Kishore V, Gao W, Tang S, Kolkman J, Gingle A, Matvienko M, Kozik A, Michelmore R, Lai Z, Rieseberg L, Knapp S (2008) SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility. TAG Theoretical and Applied Genetics 117:1021-1029

Hespeels B (2009) Contribution à l'amélioration de la résistance aux pathogènes chez la chicorée industrielle à l'aide de marqueurs moléculaires de classe 1. Mémoire présenté pour l'obtention du grade académique de master en biochimie et biologie moléculaire et cellulaire. Facultés Universitaires Notre-Dame de la Paix, Namur

Hirschhorn JN, Sklar P, Lindblad-Toh K, Lim Y-M, Ruiz-Gutierrez M, Bolk S, Langhorst B, Schaffner S, Winchester E, Lander ES (2000) SBE-TAGS: An array-based method for efficient single-nucleotide polymorphism genotyping. *Proceedings of the National Academy of Sciences* 97:12164-12169

Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T, Dong G, Sang T, Han B (2009) High-throughput genotyping by whole-genome resequencing. Cold Spring Harbor Laboratory Press

Huang X, Madan A (1999) CAP3: A DNA Sequence Assembly Program. *Genome Research* 9:868-877

ICEDD (2007) Situation environnementale des industries : l'industrie de la transformation de la betterave et de la chicorée

Ingala L, López M, Darino M, Pergolesi M, Diéguez M, Sacco F (2012) Genetic analysis of leaf rust resistance genes and associated markers in the durable resistant wheat cultivar Sinvalocho MA. *TAG Theoretical and Applied Genetics*:1-10

Iwata H, Ninomiya S (2006) AntMap : Constructing Genetic Linkage Maps Using an Ant Colony Optimization Algorithm. *Breeding science* 56:371-377

Ji X, Van den Ende W, Van Laere A, Cheng S, Bennett J (2005) Structure, Evolution, and Expression of the Two Invertase Gene Families of Rice. *Journal of Molecular Evolution* 60:615-634

Jimenez-Gomez J, Maloof J (2009) Sequence diversity in three tomato species: SNPs, markers, and molecular evolution. *BMC Plant Biology* 9:85

Katti MV, Ranjekar PK, Gupta VS (2001) Differential Distribution of Simple Sequence Repeats in Eukaryotic Genome Sequences. *Molecular Biology and Evolution* 18:1161-1167

Kaur N, Gupta AK (2002) Applications of inulin and oligofructose in health and nutrition. *J Biosci* 27:703-714

Kiers AM, Mes TH, 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

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

Koozik A, Truco M-J, Yang B, Matvienko M, Lavelle D, Hill T, Ashrafi H, Chin-Who S, Xu H, Stoffel K, Van Deynze A, Xu X, Huang S, Michelmore R (2011) The genetic map and genome sequencing of lettuce. Plant and Animal Genomes XIX Conference, San Diego

Kozik A (2007) Python SDP (Python SNP Discovery Pipeline) - CDS assembly and SNP discovery - example for *Cichorium intybus* and *Cichorium endivia* ESTs.

Kubo A, Arai Y, Nagashima S, Yoshikawa T (2004) Alteration of sugar donor specificities of plant glycosyltransferases by a single point mutation. *Archives of Biochemistry and Biophysics* 429:198-203

Kuzniar A, van Ham RC, Pongor S, Leunissen JA (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet* 24:539-551

Kwok P-Y (2001) Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics* 2:235-258

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newberg LA, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1:174-181

Lawson MJ, Zhang L (2006) Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *BioMed Central Ltd.*

Lecerf J (2009) Effets métaboliques du fructose et du miel. *Phytothérapie* 7:83-86

Leclerc L (2007) Contribution à la réalisation d'une carte génétique de la chicorée industrielle (*C. intybus*) à l'aide de marqueurs moléculaires géniques de type SSR et SNP. Rapport de stage de 2ème Master en

Sciences Biologiques. Facultés Universitaires Notre-Dame de la Paix, Namur

Legrand S, Hendriks T, Hilbert J-L, Quillet M-C (2007) Characterization of expressed sequence tags obtained by SSH during somatic embryogenesis in *Cichorium intybus* L. *BMC Plant Biology* 7:27

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* 4:203-221

Li Y-C, Korol AB, Fahima T, Nevo E (2004) Microsatellites Within Genes: Structure, Function, and Evolution. *Molecular Biology and Evolution* 21:991-1007

Lincoln S, Daly M, Lander E (1993) Constructing Genetic Linkage Maps with MAPMAKER/EXP Version 3.0. A Tutorial and Reference Manual. A Whitehead Institute for Biomedical Research Technical Report Third Edition

Liu Q, Thorland EC, Heit JA, Sommer SS (1997) Overlapping PCR for Bidirectional PCR Amplification of Specific Alleles: A Rapid One-Tube Method for Simultaneously Differentiating Homozygotes and Heterozygotes. *Genome Research* 7:389-398

Livingston D, Hinch D, Heyer A (2009) Fructan and its relationship to abiotic stress tolerance in plants. *Cellular and Molecular Life Sciences* 66:2007-2023

Livingston DP, Henson CA (1998) Apoplastic Sugars, Fructans, Fructan Exohydrolase, and Invertase in Winter Oat: Responses to Second-Phase Cold Hardening. *Plant Physiology* 116:403-408

Lorieux M (1994) Aspects statistiques de la cartographie des marqueurs moléculaires. Documents de travail de la Mission Biométrie du CIRAD Centre de coopération internationale en recherche agronomique pour le développement N°1-94

Lu X, Wang H, Liu B, Xiang J (2012) Three EST-SSR Markers Associated with QTL for the Growth of the Clam *Meretrix meretrix* Revealed by Selective Genotyping. *Marine Biotechnology*:1-10

Lynch M (2010) Evolution of the mutation rate. *Trends in Genetics* 26:345-352

Macdonald SJ, Pastinen T, Genissel A, Cornforth TW, Long AD (2005) A low-cost open-source SNP genotyping platform for association mapping applications. BioMed Central Ltd.

Madrid E, Rajesh P, Rubio J, Gil J, Millán T, Chen W (2012) Characterization and genetic analysis of an EIN4-like sequence (CaETR1) located in OTL_{AR1} implicated in ascochyta blight resistance in chickpea. Plant Cell Reports: 1-10

Manderyck B (2011) Numero special chicorée: une année chargée en essais chicorée. le Beteravier. Institut royal belge pour l'amelioration de la betterave ASBL

Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S (2006) EGAssembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Research 34:W459-W462

McConnell M, Mamidi S, Lee R, Chikara S, Rossi M, Papa R, McClean P (2010) Syntenic relationships among legumes revealed using a gene-based genetic linkage map of common bean (*Phaseolus vulgaris* L.). TAG Theoretical and Applied Genetics 121:1103-1116

McHale L, Truco M, Kozik A, Wroblewski T, Ochoa O, Lahre K, Knapp S, Michelmore R (2009) The genomic architecture of disease resistance in lettuce. TAG Theoretical and Applied Genetics 118:565-580

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KFX, Wing RA (2004) Sequence composition and genome organization of maize. Proceedings of the National Academy of Sciences of the United States of America 101:14349-14354

Moore G, Devos KM, Wang Z, Gale MD (1995) Cereal Genome Evolution: Grasses, line up and form a circle. Current Biology 5:737-739

Moreno-Hagelsieb G, Latimer K (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24:319-324

Moreno-Vazquez S, Ochoa OE, Faber N, Chao S, Jacobs JME, Maisonneuve B, Kesseli RV, Michelmore RW (2003) SNP-based codominant markers for a recessive gene conferring resistance to

corky root rot (*Rhizomonas suberifaciens*) in lettuce (*Lactuca sativa*).
Genome 46:1059-1069

Morgot-Gaudry J-F, Briat J-F (2004) La génomique en biologie végétale, INRA Editions edn

Muchero W, Ehlers J, Close T, Roberts P (2011) Genic SNP markers and legume synteny reveal candidate genes underlying QTL for *Macrophomina phaseolina* resistance and maturity in cowpea [*Vigna unguiculata* (L) Walp.]. BMC Genomics 12:8

Muys C (2004) Production et utilisation des microsatellites pour l'étude de la diversité intra- et inter-populations chez *Cichorium intybus* L. Mémoire présenté pour l'obtention du grade de licencié en Sciences biologiques. Facultés Universitaires Notre-Dame de la Paix, Namur

Nadeau JH, Sankoff D (1998) Counting on comparative maps. Trends in Genetics 14:495-501

Ooi LG, Ahmad R, Yuen KH, Liong MT (2010) *Lactobacillus acidophilus* CHO-220 and inulin reduced plasma total cholesterol and low-density lipoprotein cholesterol via alteration of lipid transporters. Journal of dairy science 93:5048-5058

Ossowski S, Schneeberger K, Lucas-Lleda JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M (2010) The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. Science 327:92-94

Pannecouque J, Waes JV, Waes CV (2011) Catalogue belge de description et de recommandation pour la chicorée industrielle. Instituut voor landbouw- en visserijonderzoek mededeling ILVO

Pollock P (1986) Fructans and the metabolism of sucrose in vascular plants. New Phytologist 104:1-24

Pool-Zobel BL, Sauer J (2007) Overview of Experimental Data on Reduction of Colorectal Cancer Risk by Inulin-Type Fructans. The Journal of nutrition 137:2580S-2584S

Quraishi U, Abrouk M, Bolot S, Pont C, Throude M, Guilhot N, Confolent C, Bortolini F, Praud S, Murigneux A, Charmet G, Salse J (2009) Genomics in cereals: from genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection. Functional & Integrative Genomics 9:473-484

Radwan O, Bouzidi MF, Mouzeyar S (2011) Molecular Characterization of Two Types of Resistance in Sunflower to *Plasmopara halstedii*, the Causal Agent of Downy Mildew. *Phytopathology* 101:970-979

Ramchiary N, Nguyen VD, Li X, Hong CP, Dhandapani V, Choi SR, Yu G, Piao ZY, Lim YP (2011) Genic Microsatellite Markers in *Brassica rapa*: Development, Characterization, Mapping, and Their Utility in Other Cultivated and Wild *Brassica* Relatives. *DNA Research* 18:305-320

Rasmussen SW (2002-2009) *SEQtools*, a software package for analysis of nucleotide and protein sequences.

Ren Y, Zhao H, Kou Q, Jiang J, Guo S, Zhang H, Hou W, Zou X, Sun H, Gong G, Levi A, Xu Y (2012) A High Resolution Genetic Map Anchoring Scaffolds of the Sequenced Watermelon Genome. *PLoS ONE* 7:e29453

Riar DS, Rustgi S, Burke IC, Gill KS, Yenish JP (2011) EST-SSR Development from 5 *Lactuca* Species and Their Use in Studying Genetic Diversity Among *L. serriola* Biotypes. *Journal of Heredity* 102:17-28

Richard G-F, Kerrest A, Dujon B (2008) Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. *Microbiology and Molecular Biology Reviews* 72:686-727

Richard GF, Paques F (2000) Mini- and microsatellite expansions: the recombination connection. *EMBO reports* 1:122-126

Ritsema T, Hernández 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. *The Plant Journal* 48:228-237

Roberfroid M, Gibson GR, Hoyles L, McCartney AL, Rastall R, Rowland I, Wolvers D, Watzl B, Szajewska H, Stahl B, Guarner F, Respondek F, Whelan K, Coxam V, Davicco M, Lacotoing L, Wittrant Y, Delzenne N, Cani P, Neyrinck A, Meheust A (2010) Prebiotic effects: metabolic and health benefits. *British Journal of Nutrition* 104:S1-S63

Rocha JR, Catana R, Ferreira BS, Cabral JMS, Fernandes P (2006) Design and characterisation of an enzyme system for inulin hydrolysis. *Food Chemistry* 95:77-82

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Press H (ed) *Bioinformatics Methods*

and Protocols: Methods in Molecular Biology Krawetz S, Misener S, Totowa

Salse Jrm (2012) In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. *Current Opinion in Plant Biology* 15:122-130

Samouelian F, Gaudin V, Boccara M (2009) *Génétique moléculaire des plantes*, Editions QUAE edn

Sawa M, Kay SA (2011) GIGANTEA directly activates Flowering Locus T in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences* 108:11698-11703

Schiex T, Gaspin C (1997) CARTHAGENE: constructing and joining maximum likelihood genetic maps. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* 5:258-267

Schranz ME, Lysak MA, Mitchell-Olds T (2006) The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends in Plant Science* 11:535-542

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nat Biotech* 18:233-234

Schwember AsR, Bradford KJ (2010) Quantitative trait loci associated with longevity of lettuce seeds under conventional and controlled deterioration storage conditions. *Journal of Experimental Botany* 61:4423-4436

Sedbrook JC, Carroll KL, Hung KF, Masson PH, Somerville CR (2002) The *Arabidopsis* SKU5 gene encodes an extracellular glycosyl phosphatidylinositol-anchored glycoprotein involved in directional root growth. *The Plant cell* 14:1635-1648

Simko I (2009) Development of EST-SSR Markers for the Study of Population Structure in Lettuce (*Lactuca sativa* L.). *Journal of Heredity* 100:256-262

Simko I, Hayes R, Truco M, Michelmore R (2011) Mapping a dominant negative mutation for triforine sensitivity in lettuce and its use as a selectable marker for detecting hybrids. *Euphytica* 182:157-166

- Sobrinho B, Brion M, Carracedo A (2005) SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Science International* 154:181-194
- Spannagl M, Mayer K, Durner J, Haberer G, Frohlich A (2011) Exploring the genomes: From Arabidopsis to crops. *Journal of Plant Physiology* 168:3-8
- Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *The Plant Journal* 3:739-744
- Tautz D (1993) Notes on the definition and nomenclature of tandemly repetitive DNA sequences. *EXS* 67:21-28
- TheGeneOntologyConsortium (2008) The Gene Ontology project in 2008. *Nucleic Acids Research* 36:D440-D444
- Timme RE, Kuehl JV, Boore JL, Jansen RK (2007) A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. *American Journal of Botany* 94:302-312
- Timms L, Jimenez R, Chase M, Lavelle D, McHale L, Kozik A, Lai Z, Heesacker A, Knapp S, Rieseberg L, Micheltore R, Kesseli R (2006) Analyses of synteny between *Arabidopsis thaliana* and species in the Asteraceae reveal a complex network of small syntenic segments and major chromosomal rearrangements. *Genetics* 173:2227-2235
- Truco M, Antonise R, Lavelle D, Ochoa O, Kozik A, Witsenboer H, Fort S, Jeuken M, Kesseli R, Lindhout P, Micheltore R, Peleman J (2007) A high-density, integrated genetic linkage map of lettuce (*Lactuca* spp.). *TAG Theoretical and Applied Genetics* 115:735-746
- Van Cutsem P, du Jardin P, Boutte C, Beauwens T, Jacquemin S, Vekemans X (2003) Distinction between cultivated and wild chicory gene pools using AFLP markers. *TAG Theoretical and Applied Genetics* 107:713-718
- Van den Ende W, Michiels A, De Roover J, Verhaert P, 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. *The Plant Journal* 24:447-456

- Van Laere A, Van Den Ende W (2002) Inulin metabolism in dicots: chicory as a model system. *Plant, Cell & Environment* 25:803-813
- Van Os H, Stam P, Visser R, Van Eck H (2005) RECORD: a novel method for ordering loci on a genetic linkage map. *TAG Theoretical and Applied Genetics* 112:30-40
- Van Stallen N (2003) Cultivar identification, genetic linkage mapping and QTL analysis in *Cichorium intybus* L. var. *Foliosum* Hegi. Proefschrift voorgedragen tot het behalen van de grad van Doctor in de Toegepaste Biologische Wetenschappen. Katholieke Universiteit Leuven
- Van Stallen N, Noten V, Demeulemeester M, Proft MD (2000) Identification of commercial chicory cultivars for hydroponic forcing and their phenetic relationships revealed by random amplified polymorphic DNAs and amplified fragment length polymorphisms. *Plant Breeding* 119:265-270
- Van Stallen N, Vandenbussche B, Verdoodt V, De Proft M (2003) Construction of a genetic linkage map for witloof (*Cichorium intybus* L. var. *foliosum* Hegi). *Plant Breeding* 122:521-525
- Varshney RK, Sigmund R, Barner A, Korzun V, Stein N, Sorrells ME, Langridge P, Graner A (2005) Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Science* 168:195-202
- Vermeulen A, Desprez B, Lancelin D, Bannerot H (1994) Relationships among *Cichorium* species and related genera as determined by analysis of mitochondrial RFLPs. *TAG Theoretical and Applied Genetics* 88:159-166
- Victoria F, da Maia L, de Oliveira A (2011) In silico comparative analysis of SSR markers in plants. *BMC Plant Biology* 11:15
- Vijn I, Smeekens S (1999) Fructan: More Than a Reserve Carbohydrate? *Plant Physiology* 120:351-360
- Vilanova S, Sargent D, Arus P, Monfort A (2008) Synteny conservation between two distantly-related Rosaceae genomes: *Prunus* (the stone fruits) and *Fragaria* (the strawberry). *BMC Plant Biology* 8:67
- Voorrips RE (2002) MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *Journal of Heredity* 93:77-78

Vos P, Hogers R, Bleeker M, Reijans M, Lee Tvd, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 23:4407-4414

Weigel D, Alvarez J, Smyth DR, Yanofsky MF, Meyerowitz EM (1992) LEAFY controls floral meristem identity in *Arabidopsis*. *Cell* 69:843-859

Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph. *PLoS Genet* 4:e1000212

Young WP, Schupp JM, Keim P (1999) DNA methylation and AFLP marker distribution in the soybean genome. *TAG Theoretical and Applied Genetics* 99:785-792

Yu JK, La Rota M, Kantety RV, Sorrells ME (2004) EST derived SSR markers for comparative mapping in wheat and rice. *Molecular Genetics and Genomics* 271:742-751