Développement d'une méthodologie d'analyse de la conservation de synténie chez les plantes

Cédric Muller

N° d'ordre : 2251

Thèse

présentée

pour obtenir

LE TITRE DE DOCTEUR
DE L'INSTITUT NATIONAL POLYTECHNIQUE DE TOULOUSE

École Doctorale :	Biologie-Santé-Biotechnologies
Spécialité :	Biosciences Végétales
Par M.	Cédric Muller

Titre de la thèse

Développement d'une méthodologie d'analyse de la conservation de synténie chez les plantes

Du génome d'*Arabidopsis* à celui du Tournesol

Soutenue le 22 Septembre 2005 devant le jury composé de :

Président	**C. Chevalet**	**Directeur du Génopôle Midi-Pyrénées, Toulouse**
Rapporteur	**P. Leroy**	**Directeur de Recherche, INRA, Clermont-Ferrand**
Rapporteur	**S. Mouzeyar**	**Chargé de Recherche, INRA, Clermont-Ferrand**
Examinateur	**T. Faraut**	**Chargé de Recherche, INRA, Toulouse**
Co-Directeur	**L. Gentzbittel**	**Professeur d'université, ENSAT, Toulouse**
Co-Directeur	**C. Brière**	**Chargé de Recherche, CNRS, Toulouse**

Résumé

Le tournesol (*Helianthus annuus* L) est l'une des principales plantes oléagineuses cultivées. L'étude des gènes sous tendant les principaux caractères agronomiques est difficile en raison de son grand génome pour lequel peu d'informations existent. Par ailleurs, les moyens financiers et techniques sont loin d'être comparables à ceux mis en place pour les céréales ou d'autres végétaux (carte physique, séquençage en masse de génome, d'ARNm, recherche des duplications, des transposons...). Pour étudier l'organisation du génome du tournesol, il a donc été envisagé une approche différente basée sur la conservation de synténie avec la plante modèle *Arabidopsis thaliana*. Les informations relatives aux gènes de la plante modèle transférées aux séquences EST et ARNm du tournesol permettent d'optimiser l'exploitation de ces séquences. Mon travail de thèse a donc consisté à mettre au point une méthode d'analyse massive du grand nombre de séquences de tournesol puis de tester expérimentalement les résultats de cette analyse afin d'obtenir de nouvelles informations sur l'organisation du génome du tournesol et d'estimer la conservation de synténie avec *Arabidopsis*.

La mise en place de la méthodologie a abouti à la création d'un serveur web appelé Iccare. Cet outil bioinformatique permet la comparaison et l'analyse d'un grand nombre de séquences de différents organismes végétaux ou animaux avec les séquences codantes des gènes de l'organisme modèle respectif, *Homo sapiens* pour les animaux et *Arabidopsis thaliana* pour les végétaux. Les résultats sont présentés sous forme graphique en combinant les informations de similitudes aux informations structurales des gènes de l'organisme modèle (introns, régions UTR). La combinaison de ces informations permet ainsi d'optimiser l'exploitation de ces séquences en utilisant les outils couplés à Iccare (définir des amorces ou des sondes). En complément de Iccare, Synteny Search (en cours d'achèvement) est un site web qui permet de rechercher les relations existantes entre les gènes d'*Arabidopsis* et du riz (conservation de synténie) ainsi que les relations de ces gènes au sein d'un même génome (ils sont uniques, dupliqués ou appartiennent à des familles multigéniques). Cet outil donne des informations complémentaires sur les gènes sélectionnés à partir d'Iccare afin de vérifier et d'interpréter les résultats expérimentaux (nombre de fragments amplifiés ou de clones BAC positifs).

Iccare a permis de sélectionner, parmi 60 200 séquences de tournesol, 20 691 séquences présentant des similitudes avec 3 635 gènes d'*Arabidopsis*. L'organisation du génome du tournesol et la conservation de synténie avec *Arabidopsis* ont été étudiées en utilisant des séquences EST de tournesol qui présentaient des similitudes avec 195 gènes localisés sur le chromosome 5 de la plante modèle. Cent cinquante neuf d'entre elles ont servi à définir des sondes Overgo

dans deux régions différentes. Le criblage d'une banque de clones BAC avec ces sondes présente une efficacité supérieure à 70%. Des amplifications PCR sur une quarantaine de clones BAC positifs ont permis de confirmer la présence des séquences EST. Ce criblage a aussi démontré que la banque utilisée était peu couvrante pour les régions étudiées ne permettant pas le regroupement en contig des clones BAC de sondes voisines chez la plante modèle. Par contre, le regroupement des clones BAC positifs à une même sonde montre qu'il existe plusieurs localisations pour cette sonde, conformément aux informations préalables fournies par Synteny Search. Parallèlement aux hybridations, 51 séquences EST ont servi de matrice à la définition de couples d'amorces spécifiques de régions conservées de part et d'autre d'introns. L'amplification par PCR sur différents cultivars de tournesol présente une efficacité de 90% et le taux de polymorphisme de taille observé et vérifié sur gel d'agarose est de 15%, ce qui a permis d'intégrer 7 nouveaux marqueurs moléculaires à une carte génétique issue de lignées recombinantes (Rachid Al-chaarani *et al.*, 2004). Ces 7 marqueurs ne sont pas liés entre eux et n'ont donc pas permis de définir de conservation de synténie entre tournesol et *Arabidopsis*.

En conclusion, Iccare permet de facilement et rapidement traiter un grand nombre de données, de transférer les informations structurelles des gènes pour faciliter et optimiser la définition d'amorces et de sondes et d'interpréter les résultats expérimentaux à l'aide de Synteny Search. En revanche, l'organisation du génome du tournesol et la synténie existant avec *Arabidopsis* n'ont pas pu être évaluées clairement. Les données préliminaires obtenues laissent à penser que l'espace entre les gènes du tournesol semble proportionnel à la différence de taille observée avec le génome d'*Arabidopsis*, expliquant ainsi la difficulté à regrouper les clones BAC en contig. Un mode d'évolution du génome du tournesol possible est donc l'augmentation de distance intergénique pouvant être due à des transposons et rétrotransposons.

Abstract

The sunflower (*Helianthus annuus* L) is one of most cultivated oil-seed plant. His great genome size makes difficult the study of gene of interest (agronomic traits) especially that few information about it are available. Others plants or crops have more technical an financial means than for sunflower (physical map, massive genome and mRNA sequencing, duplication search, transposons...). In order to study the sunflower genome organization, we have developped a new method based ont synteny conservation with the model plant *Arabidopsis thaliana*. The transfert of gene information to the EST ou mRNA sequences of the sunflower permits to optimize the sequence exploitation. My thesis works consist in developping a new methology of massive analysis of sunflower sequences then test experimentally the results in order to get new information about the genome organization of the sunflower and to estimate the degree of synteny with the *Arabidopsis* genome.

The methodology leads to the creation of a web server called Iccare. This bioinformatic tool permits to compare and analyse a high number of sequences from differents plants or animals organisms with the coding sequences of gene from the model organism which is respectively *Homo sapiens* for the animals and *Arabidopsis thaliana* for the plants. The results are displayed according to the location of the genes on the chromosomes of the reference organism. Genes structure information and sequence similarities are combined in a graphical representation in order to optimize the sequence exploitation by designing primers or probes. The Synteny Search web site (not finished yet) complete the Iccare web server, it permits to search the relation existing between the *Arabidopsis* genes and those of the rice (synteny conservation) and the relation between the genes within a genome (lonely genes, duplicated genes or genes belonging to multigenic family). This tool allows the access the information about the selected genes on Iccare in order to verify and interpret the experimental results (amplified fragment number or hybrided BAC clones).

The Iccare web server had permit to select among 60,200 sunflower sequences, 20,691 sequences presenting similarities with 3,635 *Arabidopsis* genes. The genome organization of the sunflower and the synteny conservation with *Arabidopsis* had been studied by using EST sequences of the sunflower wich present some similarities to 195 genes located on the chromosome 5 of the model plant. One hundred and fifty nine EST sequences had been used to design Overgo probes in two differents regions of the fifth chromosome. These probes have been screened on a BAC clones library and over 70% were efficient. PCR amplification on forty BAC clones had confirmed the screening. The screening had also shown that the BAC

clones library is less covering than expected in the studied regions. Some BAC clones had been fingerprinted and contig had been made showing different location for some Overgo probes according to the Synteny Search web site results. In addition to the screening, fifty one EST sequences had been used to design primer couple on conserved exons regions on both sides of an intron. PCR amplification on different sunflower cultivar with the primer couples presents 90% of efficiency and the size polymorphism rate on agarose is about 15%. Seven primer couples are used as molecular marker and are assigned on a genetic map (Recombinant Inbreed Lines). These molecular markers are not linked and any syntenic blocks can be defined between the sunflower genome and the *Arabidopsis* genome.

In conlusion, the Iccare web server permits to facilitate and accelerate the traitement of high number data, to transfert structural information of genes in order to optimize the primer design and the probe design and to interpret experimental results aided by the Synteny Search web site. In spite of these new tools, the genome organization of the sunflower and the synteny conservation with the *Arabidopsis* genome are not elucidated. The preliminary data leave with thoughts that space between the sunflower genes seems proportional to the differente genome size between the sunflower genome and the *Arabidopsis* genome. These data can explain the difficulty to contig BAC clones. One evolution of the sunflower genome possibility is the increase of intergenic length due to transposons and retrotransposons.

Remerciements

Ce manuscrit est la conclusion de quatre longues années de travail qui ont vu l'aboutissement de nombreuses choses (Matrix, le Seigneur des Anneaux, Star Wars, mais pas Harry Potter). Je profite de cette page où tout m'est permis pour **remercier toutes les personnes qui m'ont supporté (dans tous les sens du terme) au laboratoire** à commencer par :

Gilbert Alibert et Michel Petitprez, les directeurs successifs du laboratoire BAP de l'ENSAT qui m'ont accueilli. Laurent Gentzbittel, mon directeur principal de thèse, pour son soutien pour l'obtention de la bourse de thèse, ses nombreux conseils, sa disponibilité et ses encouragements durant ces quatre années de thèse. Christian Brière, mon co-directeur de thèse, pour son aide en informatique et la gestion du réseau.

Les membres du Jury, Philippe Leroy, Saïd Mouzeyar, Claude Chevalet, Thomas Faraut, Christian Brière et Laurent Gentzbittel, pour la correction de ce manuscrit et leur participation à la soutenance.

Thierry Liboz pour son aide sur pas mal de protocoles de manip. Françoise Jardinaud et Fabienne Vailleau pour nos discussions et pour mon appart' (merci Fabienne). Cecilia Tamborindeguy, Cécile Ben et Tarek Hewezi pour votre aide, votre soutien et aussi parce qu'on était doctorant en même temps. Annie Perrault, euh..., Gentzbittel depuis peu, toutes mes félicitations et merci pour nos nombreuses discussions pas toujours scientifiques je l'admets. Cathy Giovannini, Marie-Jo Tavella, Philippe Anson et Patrick Bermudes pour votre gentillesse, votre humour, votre disponibilité et votre aide précieuse pour tout ce qui concerne la vie du labo. Sylvie Devèze pour ton aide dans la recherche bibliographique (mon Dieu comment aurais-je fait sinon...). Toutes les personnes de phytopathologie.

Thomas Faraut, un grand grand merci, pour ton aide et ta patience dans ma formation " sur le tas " à la bioinformatique. J'ai eu beaucoup de plaisir à travailler avec toi.

Je finirai par les meilleurs (Yes c'est nous...), c'est-à-dire toutes les personnes qui ont partagé mon bureau (enfin notre bureau), Cécile Donnadieu Tonon (qui m'a supporté, ou plutôt subi depuis le début), Sébastien Moretti, Erwan David, Sophie Curneau, Emilie Beaumont, Gil, et les derniers arrivés ceux qui m'ont royalement " viré " j'accuse *loulou*,

alias Julien Sarry, et *chouchou*, Gérald Salin.

Je tiens aussi à remercier toutes les personnes avec qui j'ai donné des cours (TP, TD, cours) et qui m'ont fait apprécier cette partie du travail qu'est l'enseignement (Daniel Sayag, Farid Regad, Laurent Gentzbittel, Anne Bernadac, Jean Kallerhof, Thierry Huguet)

Merci aussi à tous les stagiaires qui ont passé de courts moments au labo, je leur souhaite une bonne continuation et merci pour les bons moments de détentes et de rigolades.

Un petit mot pour mon coach sportif, Damien Meyer, nouveau docteur qui m'a devancé de peu (lacheur, t'aurais pu faire un effort en dernière année pour qu'on continue le squash...).

Je finirais (ben ouais, j'ai pas encore fini, quatre ans c'est long) en remerciant tous les personnes qui m'ont supporté en dehors du labo :

Ma chérie, Géraldine, qui a fait preuve de beaucoup de compréhension, de gentillesse et d'amour durant ces quatre ans, et surtout les derniers mois de rédaction. Fabrice, Ingrid, Cécile, Julien, Estelle, Sylvain, Cyril et Nicole pour les nombreux moments de délire surtout pendant les premiers de l'an. Fabrice et Ingrid, de nouveau, pour la piscine et les nombreuses soirée de détente. Cécile et Julien pour les moments fort sympathiques de guitare, et la super correction du " phrancè selon Sédrik".

Un coucou aussi à lool, fifi, Gwen, Terminator, je vous oublie pas.

Merci aussi à mes parents et mes frères.

Merci à tous,

cette thèse est un peu le résultat de chacun
(enfin beaucoup de moi, mais un peu de vous aussi).

Ces quatre années de thèse ont été très enrichissantes et j'en garderai un bon souvenir (sauf la rédaction, c'est vraiment chi...). J'espère que mon travail servira à d'autres équipes et je souhaite bon courage à tous ceux qui considèreront mes travaux comme une porte ouverte vers de nouvelles aventures (Shreck 3, Spider-man 3, X-men 3, Harry potter...).

PS : j'ai failli oublier de remercier deux individus clés de ma thèse : mon ordinateur au bureau avec qui j'ai passé énormement de temps à le tapoter, parfois le frapper ainsi que mon ordinateur portable qui m'a permis de rédiger cette thèse sous le soleil de n'importe où.

PS du PS : Je remercie aussi le professeur Tournesol, enfin la plante Tournesol, et non c'est pas parce que je vais porter le titre de Docteur que je serai le méchant dans Tintin.

PS du PS du PS : Je remercie aussi toutes les personnes que j'ai oublié de citer, excusez-moi.

Table des matières

1 Introduction **1**

1.1 Le Tournesol . . . 1

1.1.1 Des Tournesols et des Hommes . . . 1

1.1.2 Le Tournesol, Une Plante d'Intérêt . . . 3

1.2 Revue Bibliographique . . . 11

1.2.1 Le Génome du Tournesol . . . 11

1.2.2 Evolution des Génomes et des Gènes . . . 14

1.2.3 Comparaison de Génomes et Synténie . . . 23

1.3 Objectifs du Travail de Thèse . . . 32

2 Création d'Outils Bioinformatiques **35**

2.1 Iccare : Optimiser l'Exploitation des Séquences EST . . . 35

2.1.1 Sélectionner les Séquences EST les plus Informatives . . . 35

2.1.2 Iccare, un Server Web Efficace . . . 37

2.1.3 Améliorations apportées à Iccare depuis la Publication . . . 47

2.1.4 Résultats Généraux chez les Plantes . . . 49

2.2 Iccare par la Pratique : la RuBisCo . . . 59

2.2.1 La RuBisCo, une enzyme indispensable . . . 59

2.2.2 La Rubisco selon Iccare : les régions non codantes . . . 63

2.2.3 Informations complémentaires . . . 67

2.2.4 Vérification Expérimentale . . . 68

2.2.5 Discussions et Conclusions sur la Rubisco . . . 77

2.3 Exploiter la Synténie d'autres Organismes . . . 81

2.3.1 Synteny Search, Rechercher la Synténie et les Duplications . . . 81

3 Vérification Expérimentale **86**

3.1 S'appuyer sur le Génome d'*Arabidopsis* . . . 86

3.1.1 Transférer les informations du génome d'*Arabidopsis* à celui du Tournesol 86

3.1.2 Le Génome d'*Arabidopsis* est-il Homéologues avec le Génome du Tournesol 88

3.2 Résultats expérimentaux . . . 89

3.2.1 Marqueurs Moléculaires et Cartes Génétiques . . . 89

3.2.2 Criblage de la Banque de Clones BAC . . . 94

3.3 Discussion des Résultats Expérimentaux . . . 106

3.3.1 Iccare permet de définir des Marqueurs Moléculaires Exploitables . . . 106

3.3.2 Iccare permet de définir des Sondes Overgo Efficaces 108
3.3.3 Organisation du Génome du Tournesol et Synténie 112

4 Conclusion Générale & Perspectives 116

5 Matériels et Méthodes 120
5.1 Analyse de la Rubisco 120
5.2 Criblage de la banque de clones BAC 122
5.2.1 Matériels & Méthodes 122
5.3 Tester les amorces et Définition de Cartes Génétiques 127
5.3.1 Matériels & Méthodes 127

A Erreurs de Séquençage 149

B Du Gène à la Protéine 151

C The Iccare Web Server 154

D Multalin pour Iccare 161

E Filtres Haute Densité de la Banque de Clones BAC 164

F Résultats des Hybridations des sondes Overgo 167

G Carte Génétique du Tournesol 170

H Migration sur Gel d'Agarose 4% 172

I Résultats de l'analyse de MapMaker 175

J Comparatif : Gènes Similaires d'*Arabidopsis* et Résultats Expérimentaux 186

K Extraction d'ADN Génomique 187

L La Technique de SSCP 188

M Miniprep 190

N Séquences des Sondes Overgo 191

O Définition des sondes Overgo *multi* 194

P Pattern de dépôt des clones BAC positifs 196

Q Extraction d'ADN de Clone BAC 198

R Séquences des Couples d'Amorces 199

Liste des tableaux

1.1 Teneur en acide gras (mg/litre) des huiles issues de différentes plantes oléagineuses. . 4
1.2 Surfaces cultivées de tournesol et production d'huile de tournesol en fonction des continents en 2002, source : FAO. 9

2.1 Résultat global de recherche de similitudes à partir des séquences EST et ARNm disponibles en janvier 2004. 52
2.2 Différents profils polymorphes des séquences EST du cultivar psc8 et de l'ARNm *X05079* du gène *rbcS* du tournesol. 68
2.3 Différents profils polymorphes des séquences EST du cultivar psc8, Ha300 et de l'ARNm *X05079* du gène *rbcS* du tournesol. 72
2.4 Différents profils polymorphes des séquences EST du cultivar psc8 et de l'ARNm *X05079* du gène *rbcS* du tournesol. 75
2.5 Profils polymorphes des séquences EST de 4 cultivars et de l'ARNm *X05079* du gène *rbcS* du tournesol. 75
2.6 Séquences des sondes Overgo Rub1 et Rub2 définies à partir de la séquence *X05079* du tournesol. 77

3.1 Correspondances entre les groupes de liaisons de notre analyse et ceux de la carte publiée. 94
3.2 Nombre de clones BAC positifs puis identifiés lors de l'hybridation de 149 sondes Overgo en fonction des membranes. 98
3.3 Efficacité des sondes Overgo en fonction du type de sondes. 98

E.1 Correspondence des plaques déposées sur les membranes. 166

N.1 Séquences des sondes Overgo définies à partir de séquences de tournesol similaires à des gènes d'*Arabidopsis* dans la région de la rubisco du chromosome 5. 191
N.2 Séquences des sondes Overgo définies à partir de séquences EST similaires à des gènes d'*Arabidopsis* localisés dans la région 0-7 Mbases du chromosome 5. 193

P.1 Motif de dépôt des plaques repiqués sur les membranes 11x7,5. 197

R.1 Séquences des différentes amorces définies à partir des séquences EST similaires aux gènes d'*Arabidopsis*. 200

Table des figures

1.1 Clytié et Apollon. . . . 1
1.2 Evolution de la production d'huile dans le monde pour le palmier, le soja, le colza et le tournesol, source : FAO. . . . 6
1.3 Distribution des surfaces cultivées (en MHa) dans le monde pour les 4 principales plantes oléagineuses en 2003, source : FAO. . . . 7
1.4 Evolution du rendement mondial graine/surface en t/Ha pour les 4 principales plantes oléagineuses, source : FAO. . . . 8
1.5 Evolution de la teneur mondiale en huile par graine pour les 4 principales plantes oléagineuses, ource : FAO. . . . 8
1.6 Evolution du rendement mondial huile/surface pour le palmier, le soja, le colza et le tournesol, source : FAO. . . . 8
1.7 Répartition des surfaces cultivées de tournesol dans le monde, source : FAO. . . . 9
1.8 Idiogramme des chromosomes haploïdes de *Helianthus annuus* dérivés de la métaphase d'après Schrader *et al.* (1997). . . . 14
1.9 Qu'est-ce qu'un polyploïde, d'après Leitch et Bennett (1997). . . . 15
1.10 Représentation des différents fragments dupliqués du génome d'*Arabidopsis thaliana*, source le site web TIGR. . . . 16
1.11 Représentation des événements présumés de polyploïdisation chez les angiospermes, d'après Adams et Wendel (2005). . . . 17
1.12 Mécanisme d'expansion des génomes chez les monocotylédones, d'après Feuillet et Keller (2002). . . . 19
1.13 Mécanisme menant à la contraction des génomes, d'après Feuillet et Keller (2002). . . . 19
1.14 Hybridation au stade métaphase des chromosomes du tournesol avec les sondes des rétrotransposons de type *gypsy*-like à gauche et de type *copia*-like à droite, d'après Santini *et al.* (2002). . . . 20
1.15 Evolution des relations entre gènes et terminologie. . . . 22
1.16 Cartographie génétique comparée (à gauche) et cartographie physique comparée (à droite), Barnes, 2002. . . . 24
1.17 Comparaison entre la carte génétique d'*Arabidopsis thaliana* et celle de *Brassica nigra*, Lagercrantz, 1998. . . . 26
1.18 Diagramme du 'Crop Circle' montrant les relations existantes à l'heure actuelle entre les génomes de 8 espèces appartenant à 3 sous-familles différentes, Devos, 2005. . . . 27
1.19 Comparaison d'une région orthologue d'orge, de riz, de sorgho et de blé diploïde, Ramakrishna *et al.* (2002). . . . 30

1.20 Comparaison de l'arrangements des gènes dans les régions génomiques orthogues d'*Arabidopsis*, de *Capsella* et de la tomate, Rossberg *et al.* (2001). 30
1.21 Comparaison entre la séquence du génome d'*Arabidopsis thaliana* et la carte génétique de *Capsella rubella*, Boivin *et al.* (2004). 31

2.1 Architecture de la partie analytique de Iccare. 38
2.2 Page d'accueil du site web Iccare. 40
2.3 Représentation du caryotype d'*A. thaliana*. 42
2.4 Représentation et répartition des gènes du chromosome 4 d'*A. thaliana*. 43
2.5 Page de description des résultats de similitudes pour le gène At4g03280 d'*A. thaliana*. 44
2.6 Résultats du blastn (à gauche) et du tbalstx (à droite) de la séquence CD847222 du tournesol contre l'ensemble des séquences codantes d'*A. thaliana*. 45
2.7 Représentation graphique de la structure du gène At4g03280 et de l'alignement local. 46
2.8 Arbre taxonomique des organismes eucaryotes végétaux, synthèse des données fournies par le NCBI, The Tree of Life, Museum of Paleontology, Lecointre et Le Guyader, 2001 et Guignard, 2001. 50
2.9 Représentation graphique du pourcentage des séquences EST et ARNm similaires à des séquences codantes de l'organisme modèle en fonction de l'organisme d'intérêt. . . 53
2.10 Répartition en pourcentage du nombre de séquences EST, à gauche, et ARNm, à droite, similaires à un gène de l'organisme modèle. 54
2.11 Répartition en pourcentage du nombre de gènes de la plante modèle en fonction des chromosomes. 57
2.12 Répartition en pourcentage du nombre de gènes similaires de la plante modèle en fonction du nombre d'organismes d'intérêt. 57
2.13 Représentation de la Ribulose-1,5-Bisphosphate CarboxylaseOxygénase d'après le site web "Protein Data Bank, molecule of the month (November 2000)". 60
2.14 Structure des différents types de gène *rbcS* d'après Wolter *et al* (1988). 61
2.15 Position des introns sur le gène *rbcS* At5g38420 d'*Arabidopsis*. 64
2.16 Comparaison d'une séquence EST et du gène *rbcS* At5g38420 d'*Arabidopsis*. 65
2.17 Alignement de la séquences ARNm avec la séquence génomique d'un gène *rbcS* du tournesol. 66
2.18 Alignement des séquences EST avec la séquence de l'ARNm *X05079* d'un gène *rbcS* du tournesol. 67
2.19 Positions et séquences des différentes amorces définies à partir de la séquence ARNm *X05079* du gène *rbcS* du tournesol. 69
2.20 Migration des produits d'amplification sur gel d'agarose 2%. 70
2.21 Migration des produits d'amplification sur gel SSCP. 71
2.22 Alignement des fragments séquencés de Ha300 issus de l'amplification de F3R1 (AIII) et de F0R4' (AVII). 73
2.23 Alignement des différents profils polymorphiques de 4 cultivars différents de tournesol. 76
2.24 Amplification par F0R4 des clones BAC positifs avec les sondes Rub1 et Rub2. . . . 77
2.25 Duplication du génome d*Arabidopsis*, image créée par Synteny Search. 83

2.26 Image créée par Synteny Search avec le “Content” à gauche, la “Synteny” au milieu et la “Duplication” à droite. 85

3.1 Répartition des *gènes similaires* sur les chromosomes d'*Arabidopsis*. 87
3.2 Position des gènes d'*Arabidopsis* similaires aux séquences de tournesol cartographiées. 90
3.3 Migration des produits d'amplification des amorces définies à partir de séquences EST de tournesol similaires sur 2 ou 3 cultivars de tournesol. 91
3.4 Migration des produits d'amplification des amorces définies à partir de séquences EST de tournesol similaires sur 5 cultivars de tournesol. 92
3.5 Résultats de l'hybridation de 149 sondes Overgo sur la membrane F après 5 jours et demi d'exposition à -80°C. 95
3.6 Résultats de l'hybridation de 4 sondes Overgo sur les nouvelles membranes. 96
3.7 Distribution du nombre de clones BAC positifs par sonde Overgo. 99
3.8 Duplication des gènes d'*Arabidopsis* pour la région 0-7 Mbases et 14,9-16,3 Mbases du chromosome 5 avec le reste du génome. 100
3.9 Distribution du nombre de clones BAC positifs par gène potentiel. 102
3.10 Amplification PCR sur les clones BAC positifs à 10 sondes Overgo. 104
3.11 Profil de digestion *Hind*III de clones BAC spécifiques de la sonde 39. 105

5.1 Position des gènes similaires d'*Arabidopsis* aux séquences EST ou ARNm de tournesol exploitées sur le chromosome 5 d'*Arabidopsis*. 124
5.2 Position des gènes d'*Arabidopsis* similaires aux séquences de tournesol qui servent à la définition des amorces. 128

A.1 Alignement multiple de 9 séquences EST contre 4 séquences codantes de gènes d'*Arabidopsis* qui illustrent bien les erreurs de séquençage des EST. 150

B.1 Représentation schématique de la transcription et de la traduction d'un gène. 152

D.1 Alignement multiple de dix séquences avec Multalin originale à gauche et Multalin modifiée à droite. 162

E.1 Pattern de dépôt en double dépôt 4x4 des clones BAC sur les membranes. 164

G.1 Carte génétique du croisement PAC2xRHA266 d'après Gentzbittel *et al* (1999). 170
G.2 Carte génétique du croisement PAC2xRHA266 d'après Gentzbittel *et al* (1999). 171

H.1 Migration des produits d'amplification des couples d'amorces 4 et 8 à partir des ADN des lignées recombinantes. 172
H.2 Migration des produits d'amplification des couples d'amorces 10 et 25 à partir des ADN des lignées recombinantes. 173
H.3 Migration des produits d'amplification des couples d'amorces 40 et 45 à partir des ADN des lignées recombinantes. 174

P.1 Pattern de repiquage des plaques de clones BAC positifs qui serviront aux membranes. 196

Chapitre 1

Introduction

1.1 Le Tournesol

1.1.1 Des Tournesols et des Hommes

Une légende de la mythologie grecque rapportée par Ovide dans son livre « Les Métamorphoses IV/256 » dit qu'Apollon, dieu du Soleil, avait pour maîtresse Clytié, nymphe des eaux, ainsi que sa soeur, Leucothoé. Jalouse, Clytié dénonça cette liaison à son père, Océan, qui enterra vive sa fille cadette. Apollon, trahi, abandonna Clytié qui se laissa dépérir (Figure 1.1).

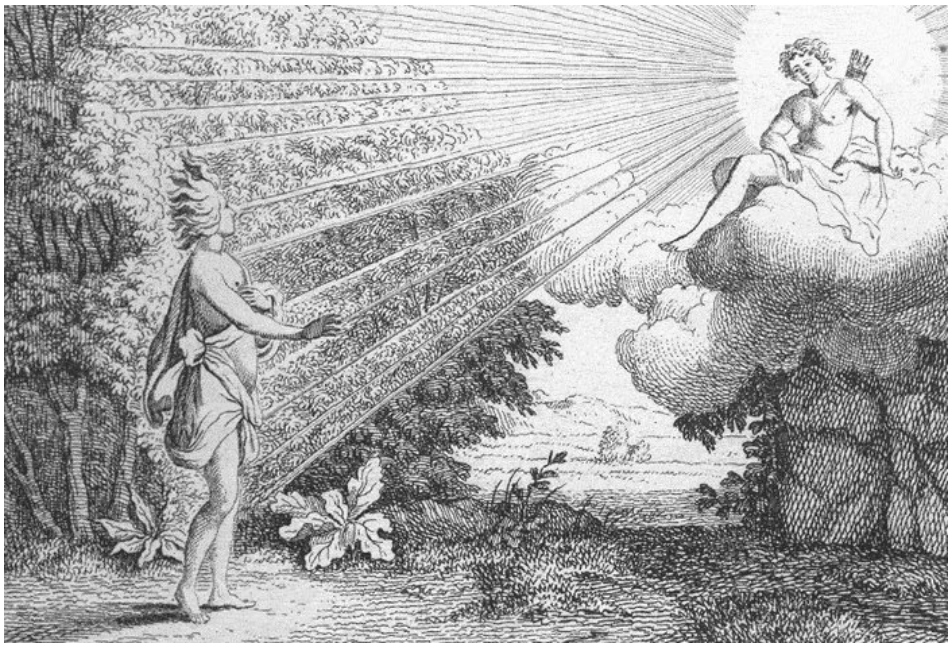


FIG. 1.1 – Clytié et Apollon.

> « Pour Clytié, bien que l'amour pût être l'excuse de son dépit, et son dépit celle de sa dénonciation, le dieu du jour ne la revit plus et désormais renonça aux joies de Vénus en sa compagnie. Depuis ce jour, la nymphe, entraînée par son amour à un acte de folie, dépérit, incapable de rien supporter ; et, nuit et jour, elle reste en plein air, assise sur la terre nue, nu-tête, échevelée. Neuf jours de suite, sans boire ni manger, elle nourrit son jeûne de simple rosée et de ses larmes, sans se lever de terre. Elle contemplait seulement la face du dieu et suivait sa course en tournant vers lui son visage. Ses membres, dit-on, adhérèrent au sol ; une pâleur livide les décolore en partie et les transforme en tiges exsangues ; une partie reste rouge et une fleur toute semblable à la violette cache sa tête. Bien que retenue par sa racine, elle tourne vers son cher Soleil et, même métamorphosée, elle lui garde son amour. »

Ainsi, Clytié fut transformée en hélianthe, condamnée à suivre la course du Soleil pour l'éternité. L'hélianthe provient du grec « helianthos » issu de *hélios* (le soleil) et de *anthos* (la fleur). Les latins, quant à eux, utilisaient le terme *helianthus*. La mythologie a sûrement été réinterprétée à la Renaissance car il n'existait pas de tournesol en Europe lorsqu'Ovide a rédigé ces textes, le terme hélianthe devait sûrement désigner une autre plante que le tournesol. Mais aujourd'hui, le tournesol porte le nom d'espèce *Helianthus annuus* : *Helianthus* pour plante qui suit le Soleil et *annuus* lui a été ajouté pour sa durée de vie annuelle. Son nom usuel français « Tournesol », quant à lui, est dérivé de sa désignation espagnole ou italienne. Les espagnols ou les italiens utilisaient le terme "Tornasole", la plante qui tourne vers le soleil, duquel le nom français est dérivé. Ce nom illustre la particularité héliotropique du tournesol, qui, dans son jeune âge, s'oriente en fonction du soleil.

Au delà de la mythologie et des origines de son nom, le tournesol a une grande place dans la culture populaire. Il apparaît dans un grand nombre de superstitions et croyances de nombreux pays. Une croyance américaine dit que si l'on grave un voeu sur le cœur d'une fleur de tournesol sur pied et si, lorsqu'elle se fane à l'automne, on peut toujours y distinguer les mots que l'on y écrit, cela signifie que le vœu se réalisera avant la fin de l'année. En Espagne, le tournesol est une fleur qui attire la chance. En Hongrie et en République Tchèque, les graines de tournesol favoriseraient la fécondation des femmes souffrant de stérilité et déposer une fleur de tournesol sur le rebord d'une fenêtre en début de grossesse permettrait la venue d'un garçon. Les chinois la considèrent comme une nourriture d'immortalité, etc.

Le tournesol est une plante qui fascine l'homme depuis la nuit des temps. Dignes repésentantes du Soleil sur Terre, les fleurs de tournesol étaient utilisées pour couronner les princesses Aztèques et ce peuple avait des représentations de fleur de tournesol en or qu'il plaçait dans leurs temples. Depuis toujours, le tournesol a beaucoup inspiré les artistes (peintres, sculpteurs, écrivains...). Il est même utilisé comme symbole, symbole d'un retour à la nature et de la culture biologique. Le tournesol a longtemps été utilisé comme plante ornementale, mais il possède aussi d'autres qualités que son esthétisme, et notamment des qualités agronomiques.

Sa culture est très ancienne et les premières traces en ont été retrouvées en Amérique du Nord. Les Indiens d'Amérique le cultivaient déjà 3000 ans avant notre ère et l'utilisaient

pour ses propriétés alimentaires, médicinales, tinctoriales et lors des cérémonies religieuses. Par la suite, il fut introduit en Europe par les Conquistadors au XVI^{e}siècle où il était principalement utilisé comme plante ornementale. C'est vers la fin du XIX^{e}siècle que les Russes cultivent le Tournesol comme plante oléagineuse. En effet, l'église orthodoxe interdisait, durant le carême, la consommation de corps gras et l'huile de Tournesol ne figurait pas dans la liste des aliments interdits. Les Russes ont alors commencé à établir des programmes de sélection et d'amélioration des variétés de tournesol afin d'accroître les rendements et ont réussi à atteindre une teneur en huile de 44% grâce à ces travaux.

De nombreux programmes de sélection ont ensuite été mis en place pour accroître le rendement du tournesol, de nouvelles lignées ont ainsi été obtenues ainsi que la création de variétés hybrides qui bénéficient d'un effet hétérosis (un accroissement des rendements comparés aux parents). Cependant, la création de variétés hybrides est difficile car elle se heurte à des phénomènes d'incompatibilités entre lignées de tournesol. C'est à partir des années 1970 que ce problème a pu être surmonté grâce à la découverte de la "Cytoplasmic Male Sterility" (CMS) ou stérilité mâle cytoplasmique par Leclercq en 1969. La CMS permet de simplifier la fabrication des variétés hybrides. Depuis cette période et jusqu'à aujourd'hui, le tournesol est cultivé et amélioré dans le monde afin d'accroître ses caractères agronomiques car en plus de la teneur en huile, la domestication a permis de sélectionner des tournesols non ramifiés (mono-capitule) dont la maturité est groupée et les graines non déhiscentes. La vigueur hybride et la stérilité mâle cytoplasmique (Leclercq, 1970) ainsi que ces différents caractères agronomiques ont permis au tournesol de devenir une plante de grande culture dans le monde entier.

1.1.2 Le Tournesol, Une Plante d'Intérêt

Produits dérivés des plantes oléagineuses

Le tournesol est une plante oléagineuse, comme le colza, le soja, l'arachide ou le palmier. Il est donc cultivé pour ses graines riches en huile mais aussi pour ces tourteaux riches en protéines végétales. Les plantes oléagineuses permettent la fabrication des huiles végétales en utilisant des méthodes d'extraction physique (par pression) et/ou par méthode chimique (par extraction avec solvant). Ces méthodes de trituration (extraction de l'huile) entraînent la formation de résidus secondaires solides appelés tourteaux qui sont riches en protéines et sont utilisés en alimentation animale. Ils constituent la 2^{e}classe d'aliments la plus importante après les céréales. Mais le produit principal de fabrication à partir des graines ou fruits d'oléagineux est l'huile.

Au cours du siècle dernier, les huiles végétales ont pris une part plus importante dans l'alimentation humaine, et ce, grâce aux modifications alimentaires concernant les matières grasses. L'apport dans l'alimentation de matière grasse était principalement d'origine animale. Ces graisses animales sont majoritairement constituées d'acides gras saturés (AGS) et

de cholestérol. Les AGS ainsi que le cholestérol, en grande quantité, ont des effets néfastes sur la santé. Des études ont montré que les AGS favorisent considérablement les troubles cardio-vasculaires et les dysfonctionnements cellulaires. En plus, ils augmentent la présence de cholestérol LDL (Low Density Lipoprotein, le "mauvais" cholestérol) au détriment du cholestérol HDL (High Density Lipoprotein, le "bon" cholestérol). Ce déséquilibre entre HDL et LDL augmente le taux de cholestérol sanguin et favorise ainsi le processus d'athérosclérose (formation de caillots dans les vaisseaux sanguins).

Les graisses animales ont donc petit à petit été remplacées par des huiles végétales. Celles-ci ont pris une part de plus en plus importante dans l'alimentation car leur composition est différente de celle des graisses animales. Les huiles végétales sont riches en acides gras mono-insaturés (AGMI) et poly-insaturés (AGPI) et contiennent en plus des phytostérols, composés de la famille des stérols (comme le cholestérol) aux propriétés protectrices.

Dans les AGMI, on retrouve l'acide oléique qui joue un rôle protecteur contre l'athérosclérose ; il abaisse le cholestérol total et le cholestérol LDL sans réduire le cholestérol HDL. Dans les AGPI, l'acide linoléique (ou Omega-6) et l'acide alpha linolénique (ou Omega-3) sont tous deux des acides gras essentiels car ils ne peuvent être synthétisés par l'organisme, ils doivent donc nécessairement être fournis par l'alimentation. Les AGPI ont des rôles essentiels au sein de l'organisme. Ils interviennent dans la fluidité membranaire du fait du repliement de la molécule à chaque double liaison, entrent dans la composition du tissu nerveux, de la rétine, participent à la croissance, à la cicatrisation de la peau, à l'agrégabilité plaquettaire (les cellules, plus souples, s'agglutinent moins) et protègeraient contre certains cancers (source : http://fderad.club.fr/lipides.htm).

Les huiles végétales sont donc très vivement conseillées dans l'alimentation. Mais toutes les huiles végétales n'ont pas les mêmes teneurs en acides gras. Certaines sont plus riches que d'autres en un ou plusieurs acides gras particuliers (Tableau 1.1) ou peuvent contenir d'autres éléments essentiels comme les vitamines (le tournesol est riche en vitamine E, alpha-tocophérone). Ces teneurs différentes en acides gras font qu'il est conseillé d'alterner plusieurs huiles végétales afin de satisfaire les besoins essentiels.

Aliments	**AGS**	**AGMI**	**AGPI**	
			linoléique	linolénique
Huile de palme	21-51	39-59	10-19	0,3-0,9
Huile d'arachide	20-22	41-60	20	<0,1
Huile de soja	15	22	56	7
Huile de tournesol	11	22	67	0,1
Huile de colza	8	61	22	9

TAB. 1.1 – Teneur en acide gras (mg/litre) des huiles issues de différentes plantes oléagineuses.

La grande majorité des huiles végétales sert à l'alimentation humaine, environ 70% de la production mondiale en 2002. Le reste est utilisé dans l'industrie (environ 28%) et les 2% restant servent à la préparation de sauces et d'assaisonnement ou à l'alimentation animale. L'industrie utilise les huiles végétales comme solvant ou lubrifiant (savons, détergents, peintures, résines glycérophtaliques, lubrifiants, cosmétiques, encre...), mais depuis quelques

années se développe une nouvelle utilisation des huiles végétales : les Biocarburants.

Ces biocarburants viennent tous de la transformation des plantes en énergie liquide. Les plantes riches en sucre comme la betterave, la pomme de terre, le maïs, la canne à sucre, peuvent être transformées en alcools. Le produit ainsi obtenu, appelé éthanol, peut être utilisé à 100% en remplacement de l'essence. A partir de plantes oléagineuses (colza, tournesol, ricin, palme ou noix de coco), deux types de produits peuvent être obtenus : des Huiles Végétales Pures (HVP) ou Brutes (HVB) et un produit plus élaboré appelé Ester Méthylique d'Huile Végétale (EMHV). Les esters sont issus de la fluidification de l'huile après environ 25 opérations industrielles pour obtenir ce que certains appellent le Biogazole (ou Diester ou Dieselbi). Les Huiles Végétales (HVB ou HVP) sont le résultat direct du pressurage des graines d'oléagineux (colza, tournesol) puis d'une filtration, elles demandent beaucoup moins de préparation que les esters et les alcools. Les HVP et les EMHV viennent en remplacement du gazole (sans avoir à modifier les moteurs diesel existants).

Parmi ces 3 types de biocarburants, le plus rentable (en terme de consommation d'énergie à la production) est le HVB. La transformation industrielle pour produire des éthanols consomme beaucoup d'énergie, et il faut l'équivalent énergétique de 900 litres d'éthanol pour en fabriquer 1 000 litres ; de même pour les diesters, pour obtenir 1 000 litres d'ester, il faut en consommer l'équivalent énergétique de 680 litres ; alors que pour les HVB, l'équivalent énergétique de 300 litres d'huile (colza ou tournesol) permet d'obtenir 1 000 litres de carburant : on obtient donc deux fois plus de carburant par hectare de culture en fabriquant de l'huile plutôt que de l'ester, et six fois plus par rapport à l'éthanol (source : http://www.biorespect.com/lesnews.asp?ID=4\&NEWSID=70).

Les huiles végétales sont devenues incontournables dans notre alimentation durant le siècle dernier, bientôt elles devraient aussi l'être pour nos voitures et les tourteaux sont devenus indispensables en alimentation animale surtout depuis le retrait des farines animales. Les huiles végétales ont un bel avenir devant elles, autant dans l'alimentaire que sur le nouveau marché des biocarburants et avec les tourteaux qui ne sont pourtant que les produits secondaires de l'extraction de l'huile. Parmi les plantes oléagineuses, le tournesol est une plante qui a du potentiel.

Le Tournesol face à ses Concurrents

Le tournesol est l'une des principales plantes oléagineuses cultivées pour la production d'huiles végétales avec le palmier, le colza et le soja. Il est cultivé pour sa forte teneur en huile (de 44% à 90% du poids de la graine) mais aussi pour sa teneur en protéine (pouvant atteindre jusqu'à 25% du poids de la graine) essentiel dans la production de tourteaux.

Production mondiale : en 2002, les 101 millions de tonnes (Mt) d'huile végétale produite dans le monde sont à 75% fournies par 4 plantes (soja, palmier, colza et tournesol) et 9 plantes se répartissent les derniers 25%. Le soja et le palmier sont les 2 principales plantes cultivées avec respectivement 30,1 et 25,5 Mt (soit 30 et 25% de la production mondiale)

puis le colza et le tournesol avec 12,4 et 8,1 Mt. Au cours des quarante dernières années, ces 4 plantes ont connu une forte augmentation de leur exploitation (Figure 1.2) ce qui a permis de quasiment quintupler la production mondiale d'huile végétale passant de 19 Mt en 1961 à 101 Mt en 2002 (source : Food and Agriculture Organization of the United Nations Statistical Databases, http://apps.fao.org/). Cette augmentation de la production d'huile est principalement due à l'augmentation des surfaces cultivées ainsi qu'à l'augmentation des rendements des 4 principales plantes oléagineuses.

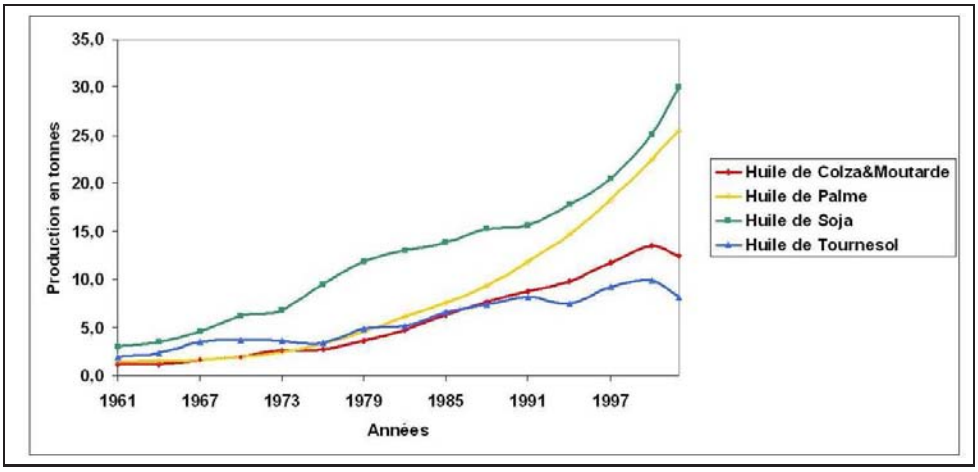


FIG. 1.2 – Evolution de la production d'huile dans le monde pour le palmier, le soja, le colza et le tournesol, source : FAO.

Surfaces cultivées : elles ont été multipliées par 3,5 au cours des quarante dernières années dans le monde. Le cumul des surfaces cultivées est passé de 40,4 millions d'hectares (MHa) en 1961 à 141,6 MHa en 2003 (source : http://apps.fao.org/). Durant cette période, la proportion relative des surfaces cultivées entre les 4 plantes s'est toujours maintenue : le soja a toujours été la principale plante cultivée et représente entre 55 et 64% des surfaces cultivées, le colza entre 14,5 et 20%, le palmier entre 5 et 9% et le tournesol entre 15,5 et 18%. Cependant, leurs répartitions géographique et d'espèce ne sont pas équitables (Figure 1.3). Ainsi, l'Amérique cultive à elle seule la moitié des surfaces mondiales d'oléagineux avec, en 2003, 72,9 MHa (cumul de l'Amérique du Nord et du Sud) dont 63,5 MHa dédiés au soja (soit 88% des surfaces cultivées en Amérique). Ensuite, vient l'Asie avec 41,5 MHa (22% des surfaces mondiales) dont la principale culture est aussi le soja avec 17,5 MHa (soit 42% des surfaces cultivées en Asie), puis l'Europe avec 19,2 MHa (13,5% des surfaces mondiales) dont la principale culture est le tournesol avec 13,5 MHa (soit 70% des surfaces cultivées en Europe), puis l'Afrique avec 6,3 MHa (4% des surfaces mondiales) dont la principale culture est le palmier avec 4,3 MHa (soit 68% des surfaces cultivées en Afrique) et enfin l'Océanie avec 1,1 MHa (moins de 1% des surfaces mondiales) dont la principale culture est le colza avec 1 MHa (soit 90% des surfaces cultivées en Océanie).

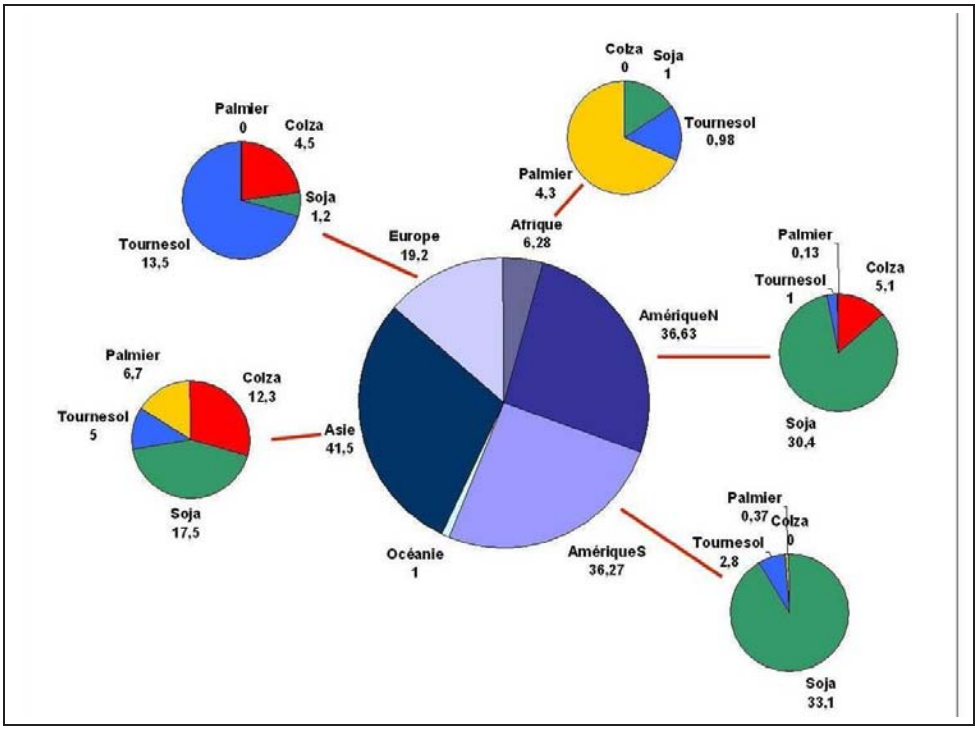


FIG. 1.3 – Distribution des surfaces cultivées (en MHa) dans le monde pour les 4 principales plantes oléagineuses en 2003, source : FAO.

Rendements : les rendements (quantité de graines par surfaces cultivées et quantité d'huile par graine) des 4 principales plantes oléagineuses ont eux aussi progressé durant les quarante dernières années. C'est le palmier qui en a le plus profité avec une augmentation du rendement quantité de graines par surface (ou fruit/surface) multipliée par 3,2 passant de 37,7 q/Ha en 1961 à 120,1 q/Ha en 2002 (Figure 1.4). Le colza et le soja ont eu une augmentation de leur rendement un peu moindre, celui-ci a été respectivement multiplié par 2,6 et 2, passant de 5,7 q/Ha à 14,8 q/Ha pour le colza et de 11,3 q/Ha à 22,9 q/Ha pour le soja. Le tournesol, quant à lui, a eu la plus faible augmentation de rendement, par 1,2, passant de 10,2 q/Ha à 12,6 q/Ha.

Le rendement de la teneur en huile (huile/graine) a peu augmenté pour les 4 plantes (Figure 1.5). L'augmentation du rendement est compris entre 1,2 et 1,7. Ce sont le colza et le tournesol qui présentent les meilleurs rendements avec respectivement 0,36 tonnes d'huile par tonne de graines (th/tg) et 0,33 th/tg (palmier et soja respectivement : 0,19 et 0,17 th/tg).

Il en résulte que la combinaison des 2 rendements fait que le palmier est la plante qui a connu la plus forte augmentation du rendement quantité d'huile par surface cultivée, avec une multiplication par 5,6 entre 1961 et 2002, passant de 4 q/Ha à 22,6 q/Ha (Figure 1.6). Le colza et le soja ont eu une augmentation du rendement similaire (par 3), passant respectivement de 1,8 à 5,5 q/Ha et de 1,3 à 3,8 q/Ha. Le tournesol, quant à lui, a eu la plus faible augmentation (par 1,4) passant de 2,9 à 4,1 q/Ha. Ces hausses du rendement sont le résultat des progrès des biotechnologies qui ont permis l'amélioration des variétés cultivées pour avoir un meilleur rendement avec moins de surfaces cultivées.

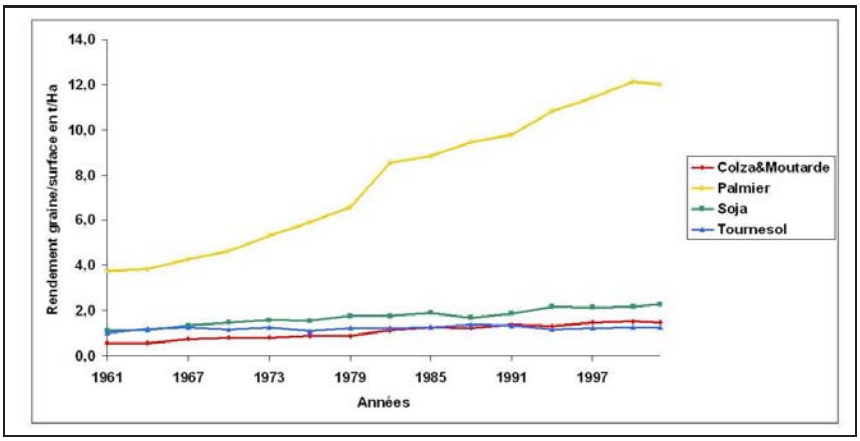


FIG. 1.4 – Evolution du rendement mondial graine/surface en t/Ha pour les 4 principales plantes oléagineuses, source : FAO.

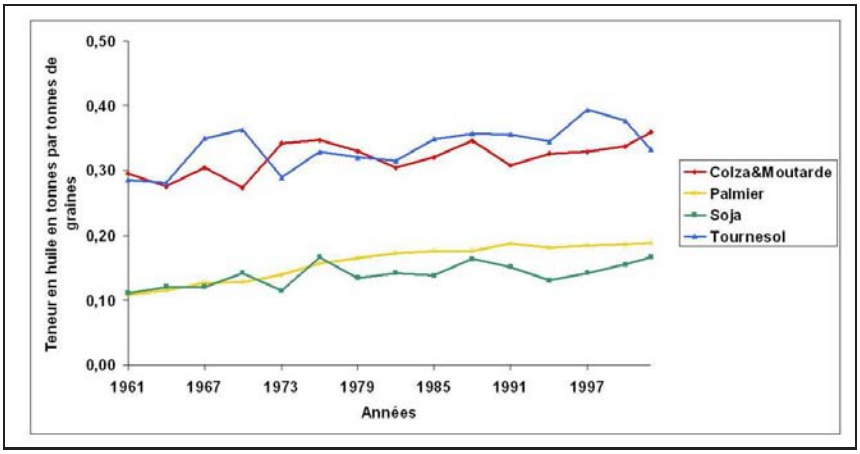


FIG. 1.5 – Evolution de la teneur mondiale en huile par graine pour les 4 principales plantes oléagineuses, ource : FAO.

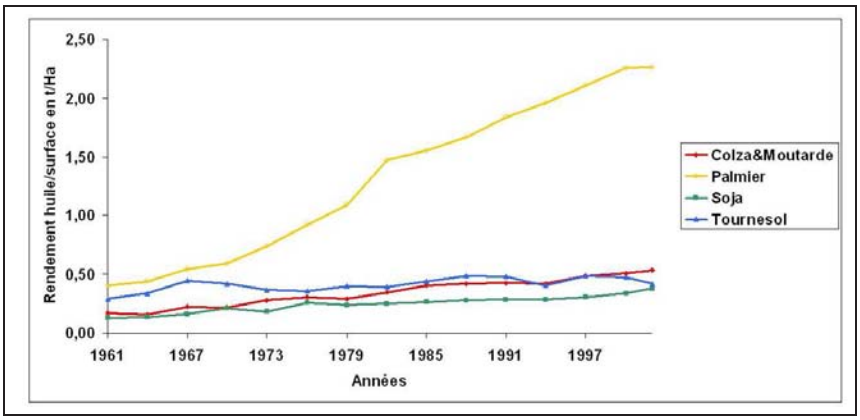


FIG. 1.6 – Evolution du rendement mondial huile/surface pour le palmier, le soja, le colza et le tournesol, source : FAO.

Le tournesol présente l'augmentation de rendement la plus faible, mais l'augmentation des surfaces cultivées a suivi celle des autres plantes ce qui fait que le tournesol reste toujours une plante majeure dans la production d'huile végétale. Mais si l'augmentation du rendement reste faible dans les années à venir celui-ci risque d'être un handicap, et les autres plantes risquent d'être privilégiées pour l'obtention d'huile végétale.

Le Tournesol : Une culture en déclin ou émergente ?

En quarante ans, les surfaces de tournesol cultivées dans le monde ont été multipliées par 3,5 passant de 6,67 MHa en 1961 à 23,3 MHa en 2003. Les principaux pays qui cultivent du tournesol sont des pays d'Europe, zone géographique (Tableau 1.2).

Continents	**Surfaces cultivées (MHa)**	**% monde**	**Production d'huile (Mt)**	**% monde**	**Rendement huile/surface (q/Ha)**
Europe	10,48	54	4,45	57	4,2
Asie	4,49	23	1,51	19	3,3
Amérique Sud	2,42	12	1,43	18	5,9
Amérique Nord	0,98	5	0,34	4	3,4
Afrique	0,94	5	0,39	5	4,1

TAB. 1.2 – Surfaces cultivées de tournesol et production d'huile de tournesol en fonction des continents en 2002, source : FAO.

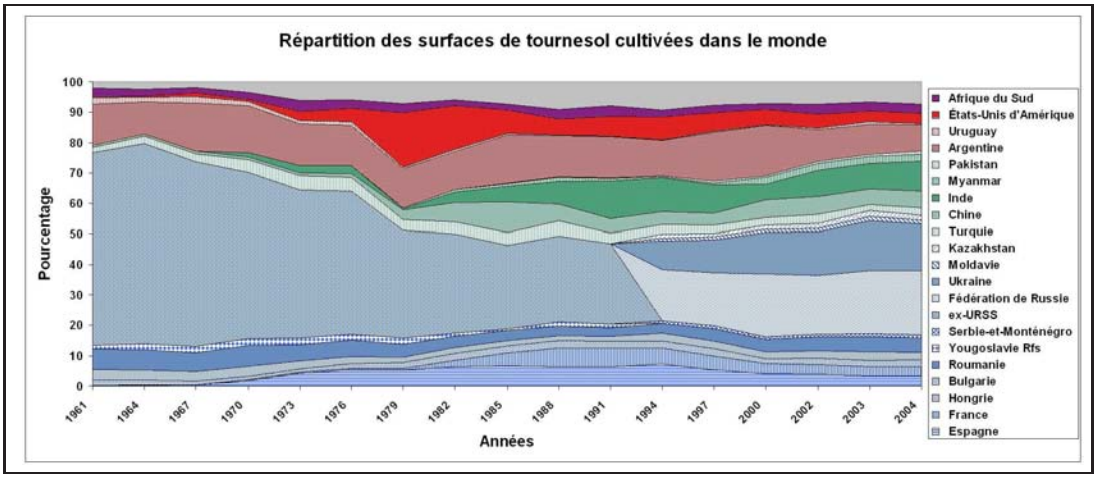


FIG. 1.7 – Répartition des surfaces cultivées de tournesol dans le monde, source : FAO.

Le nombre de pays impliqués dans la production du tournesol a énormément augmenté en quarante ans (Figure 1.7). Pour obtenir 90% des surfaces, il suffisait de compter 6 pays en 1961, en 2004, il en faut plus de 25. Aujourd'hui, de nombreux pays cultivent du tournesol, et la tendance est à la diminution des surfaces de culture pour les pays traditionnellement producteurs (Fédération de Russie et pays d'Europe de l'Ouest), au maintien pour les pays

d'Amérique (à l'exception des Etats–Unis d'Amérique) et surtout à l'augmentation des surfaces des pays d'Europe de l'Est et de l'Asie.

Pendant quarante ans, les surfaces de tournesol dans le monde ont énormément augmenté, mais les rendements en graines ou de la teneur en huile, eux, ne semblent avoir que très peu progressé pendant cette même période.

Le rendement moyen de la teneur en huile est resté très stable variant de 0,29 à 0,39 tonne d'huile par tonne de graines et le rendement moyen (quantité de graines/surface) a très peu augmenté passant de 10 q/Ha à 12,6 q/Ha. Pourtant, certains pays malgré la diminution des surfaces cultivées n'ont pas diminué leur production d'huile végétale. Ces pays ont compensé leur diminution de surface par une augmentation des rendements.

Même si ces rendements ne semblent pas avoir énormément bougé durant quarante ans au niveau mondial, il en est totalement différent au niveau des rendements par pays. Le rendement de la teneur en huile n'a effectivement que peu augmenté quel que soit le pays, par contre le rendement quantité de graines/surface est très nettement différent selon les pays. Ainsi, certains pays ont un rendement quantité graines/surface plus important que d'autres (en 2004, France : 23,6 q/MHa, Hongrie : 23,4 q/MHa, Italie : 22,3 q/MHa, Myanmar : 5,8 q/MHa, Maroc : 5,3 q/MHa, Inde : 6,0 q/MHa). De plus, l'évolution de ces rendements est très différente d'un pays à l'autre depuis quarante ans. Certains ont des rendements qui ont stagné (Maroc, Inde, Bulgarie, Italie, etc.) alors que d'autres ont pu profiter des programmes de sélection et d'amélioration des caractères agronomiques du tournesol leur permettant ainsi d'augmenter le rendement de la plante (Pakistan, Myanmar, Uruguay, Argentine, Hongrie, France). Ces programmes de sélection ont permis, notamment en France, de faire progresser les rendements de cette plante et faire qu'elle soit performante dans la production d'huile (augmentation du nombre de graine, de la qualité de l'huile, résistance aux maladies, etc.).

Au cours des quarante dernières années, le tournesol est une culture délaissée dans certains pays alors qu'elle émerge dans d'autres. Il en résulte qu'au final, le tournesol est une culture qui se maintient au niveau international. En quarante ans, le nombre de pays, les surfaces et les rendements pour la culture du tournesol ont augmenté. De nombreuses évolutions ont pu être observées au niveau des répartitions des surfaces ainsi que des rendements dans le monde. Mais en fin de compte, le tournesol reste une plante majeure dans la production d'huile végétale et de tourteaux. Pourtant cette plante est loin d'avoir donné son maximum, elle nécessite encore de nombreuses améliorations et notamment dans sa teneur en huile et sa quantité de graines car c'est une culture dont la sélection est récente (moins d'un siècle).

1.2 Revue Bibliographique

1.2.1 Le Génome du Tournesol

Depuis environ un siècle, le travail de sélection a permis d'adapter le tournesol à nos conditions de sols et de climat : longueur du cycle de la plante, résistance aux maladies, hauteur, régularité de maturation, tenue sur tige et potentiel de production. L'étude de son génome est encore plus récente et a débuté par des analyses génétiques.

Différentes analyses portant sur différents gènes intervenant dans des processus physiologiques ont été réalisées sur le tournesol, mais ces analyses sont très récentes comparées à d'autres plantes. La sélection assistée par marqueurs moléculaires n'est vieille que d'une décennie, mais ces derniers ont permis de rendre la sélection plus précise et plus rapide qu'une sélection classique. Ils ont permis la construction de cartes génétiques et l'étude des caractères agronomiques d'intérêt associés à différentes régions du génome (analyse de QTL, Quantitative Traits Loci). Plusieurs cartes génétiques ont ainsi été construites à partir de différentes lignées de tournesols et suivant différentes techniques.

Les premières cartes génétiques de tournesol ont été faites en utilisant la technique de Restriction Fragment Length Polymorphism (RFLP) ou polymorphisme des longueurs de fragments de restriction (Gentzbittel *et al.*, 1995, 1999, Jan *et al.*, 1998), puis en utilisant la technique de Random Amplified Polymorphism DNA, RAPD ou polymorphisme d'ADN par amplification aléatoire (Rieseberg *et al.*, 1995), ensuite en combinant les techniques de RFLP et d'AFLP, Amplified Fragments Length Polymorphism ou polymorphisme de longueur de fragments amplifiés (Gedil *et al.*, 2001) ou en combinant l'AFLP et les microsatellites, Simples Sequences Repeats (Mokrani *et al.*, 2002, Rachid Al-Chaarani, *et al.*, 2002) ; d'autres cartes génétiques ont été faites en utilisant uniquement des microsatellites (Yu *et al.*, 2002, Tang *et al.*, 2002) et Langar *et al.* (2003) ont combiné les AFLP à un nouveau type de marqueurs les DALP, Direct Amplification of Length Polymorphism ou amplification directe de polymorphisme de taille, pour construire la leur.

A partir de ces cartes génétiques, quelques caractères agronomiques ont été analysés comme des caractères intervenant dans l'embryogénèse somatique (Flores Berrios *et al.*, 2000), la photosynthèse (Hervé *et al.*, 2001), des processus physiologiques ou le remplissage du grain (Mokrani *et al.*, 2002, Bert *et al.*, 2003, Rachid Al-Chaarani *et al.*, 2004). Mais ce sont principalement les QTL des gènes de résistance qui ont le plus été étudiés. Parmi les principales maladies qui touchent le tournesol, des analyses QTL ont été faites sur *Sclerotinia sclerotiorum* (Bert *et al.*, 2004, Micic *et al.*, 2004, 2005, 2005), sur *Phoma macdonaldii* (Bert *et al.*, 2004, Rachid Al-Chaarani *et al.*, 2002) et plus particulièrement sur le champignon responsable du mildiou *Plasmopara halstedii*.

Le mildiou, causé par *Plasmopara halstedii*, est une des principales maladies du tournesol cultivé. Les gènes conférant la résistance à cette maladie, notés *Pl*, répondent à l'interaction gène-pour-gène formulée par Flor (1971). Les analyses QTL ont permis d'identifer sur les

cartes génétiques la présence de deux clusters de gènes de résistances aux différentes races de mildiou. Le premier cluster contient les gènes *Pl1*, *Pl2* et *Pl6* (Mouzeyar *et al.*, 1995, Vear *et al.*, 1997) et le deuxième cluster contient les gènes *Pl5* et *Pl8* (Bert *et al.*, 2002, Radwan *et al.*, 2003). Ils sont respectivement localisés sur le groupe de liaison 1 et 6 de la carte génétique de Gentzbittel *et al* (1999). Le premier cluster de gènes contient 13 gènes de type TIR-NBS-LRR, qui contiennent trois domaines de type Toll/Interleukin-1 Receptor, Nucleic Binding Site et Leucin-Rich Repeats (Gentzbittel *et al.*, 1998, Bouzidi *et al.*, 2002) alors que le deuxième cluster contient 16 gènes de type CC-NBS-LRR, qui contiennent trois domaines de type Coiled-Coil, Nucleic Binding Site et Leucin-Rich Repeat (Gedil *et al.*, 2001, Radwan *et al.* 2004, 2005).

Cette organisation des gènes de résistance en deux clusters localisés dans deux régions différentes du génome pourrait être le résultat d'une duplication puis d'une spéciation des gènes en deux types différents. Les gènes de types NBS-LRR qui possèdent le domaine TIR sont absents chez les monocotylédones (Meyers *et al.*, 1999). La duplication ou spéciation se serait donc produite après la séparation entre dicotylédones et monocotylédones.

D'autres indices laissent à penser que le génome du tournesol a subi des duplications. Sur 145 marqueurs moléculaires utilisés par Gentzbittel *et al.* (1995, 1999) pour la construction de cartes génétiques, 45 d'entre eux ont montré qu'ils étaient liés à plusieurs groupes de liaison. Ainsi sur la carte génétique de Gentzbittel *et al.* de 1999, 5 marqueurs du groupe de liaison 3 sont aussi localisés sur le groupe de liaison 8 qui lui même contient aussi 2 autres marqueurs localisés sur le groupe de liaison 2 (Figure **??**). De même, Jan *et al.* (1998) ont utilisé 232 sondes faites à partir de séquences d'ADNc pour construire une carte génétique et parmi ces sondes 32 ont une localisation multiple. Tous ces indices laissent à penser que le génome du tournesol est dupliqué (ou en partie).

D'autres analyses ont été menées sur l'ensemble du génome du tournesol. Le tournesol (*Helianthus annuus* L) est une plante composée et fait partie de la famille des *Asteraceae*, de la tribu des *Heliantheae*, de la sous-tribu des *Helianthinae* et du genre *Helianthus* (Schilling, 1997). Le genre *Helianthus* est divisé en 4 sections. La section *Helianthus* contient 11 espèces toutes diploïdes $2n = 34$ (dont le tournesol) et la section *Agrestes* contient l'espèce diploïde *Helianthus agrestis*. Ces deux sections contiennent uniquement des espèces annuelles. Les sections *Ciliares* et *Atrorubentes* contiennent respectivement 6 et 31 espèces et toutes sont pérennes.

Sossey-Alaoui *et al.* (1998) ont utilisé des marqueurs moléculaires sur 40 espèces de 3 sections du genre *Helianthus* (*Helianthus*, *Ciliares* et *Atrorubentes*) pour analyser le génome de ces plantes. Le résultat des analyses faites sur ces 3 sections du genre *Helianthus* démontre l'existence de génomes communs et spécifiques au sein des espèces des différentes sections. Ainsi 33 marqueurs sont communs à toutes les espèces, 29 sont spécifiques des espèces *Helianthus*, 56 sont spécifiques des plantes pérennes et parmi ces 56 marqueurs 24 sont spécifiques des espèces de la section *Atrorubentes*. Un génome commun a donc été mis en évidence ainsi qu'un génome spécifique de chaque section. Les espèces *Helianthus* possèdent donc le génome *C*

(commun) ainsi que le génome *H* spécifique des *Helianthus*, les espèces pérennes (*Ciliares* et *Atrorubentes*) possèdent le génome *C* ainsi que le génome *P* spécifique des plantes pérennes, et les espèces *Atrorubentes* possèdent en plus un génome *A* qui leur est spécifique.

Ces inter-relations entre génomes des espèces du genre *helianthus* sont peut être la conséquence d'un événement d'allotétraploïdisation qu'aurait subi le tournesol (et peut être d'autres espèces du même genre) entre une espèce du genre voisin *Viguiera* ($n = 8$) et une espèce d'*Helianthus* inconnue ($n = 9$) et qui serait responsable du nombre actuel de chromosome du tournesol (Heiser et Smith, 1965). Burke *et al.* (2004) ont aussi démontré l'existence de plusieurs remaniements chromosomiques (8 translocations et 3 invertions) entre deux espèces d'*Helianthus* (*Helianthus annuus* et *Helianthus petiolaris*), démontrant ainsi la dynamique du génome des plantes du genre *Helianthus*.

En fait, peu d'informations sont disponibles sur le génome du tournesol et son organisation. Schrader *et al.* (1997) ont montré que le caryotype du tournesol était constitué de 17 paires de chromosomes dont 4 paires sont acrocentriques et 13 sont meta- à submetacentriques (Figure 1.8). La taille des chromosomes varie peu entre chromosomes (entre 4,8 et 7,2 μm au stade métaphase). La taille du génome du tournesol est estimée à 3 000 Mbases, ce qui fait que la taille des chromosomes serait comprise entre 143 et 214 Mbases si on respecte les résultats du caryotype. La carte génétique la plus dense et la plus grande est proposée par Rachid Al-Chaarani *et al.* (2004) et mesure 2915,9 cM. Elle contient 21 groupes de liaisons dont les tailles varient de 329 à 20 cM. Le nombre de groupes de liaisons est supérieur au nombre de chromosomes et les groupes de liaisons devraient aussi avoir des tailles relativement proches car la taille des chromosomes varie peu. Il est donc clair que certaines régions du génome ne sont pas couvertes par des marqueurs moléculaires et que la carte génétique du tournesol n'est pas encore saturée.

Le Tournesol est une plante qui possède un grand et complexe génome. Pour valoriser et faciliter la sélection de variétés ou d'hybrides performants et résistants (ou tolérants) aux maladies, il faut avoir de plus amples informations sur le génome de cette plante et notamment accéder aux différents gènes qui participent à ces processus physiologiques (gènes de résistance mais aussi ceux qui participent aux caractères agronomiques). Bien entendu, l'idéal serait d'avoir directement accès à la séquence complète du génome du tournesol (ce qui est en fait l'objectif final à long terme, voire très long terme). Malheureusement, le tournesol possède un grand génome équivalent à la taille du génome humain. Cette taille de génome rend donc impossible la mise en place d'un programme de séquençage comme celui réalisé sur la plante modèle *Arabidopsis thaliana*, qui a une taille de génome de 125 Mbases, ou comme pour le génome humain car il faudrait mettre en place des consortiums internationaux ainsi que les mêmes moyens financiers, techniques et humains pour aboutir en moins de 10 ans à l'obtention de la séquence complète. La mise en place d'un programme de séquençage de la totalité du génome du tournesol n'est donc pas envisageable.

Puisque le séquençage du génome n'est pas envisageable, pourquoi ne pas s'intéresser qu'aux régions du génome qui contiennent les gènes. Seuls les gènes impliqués dans des

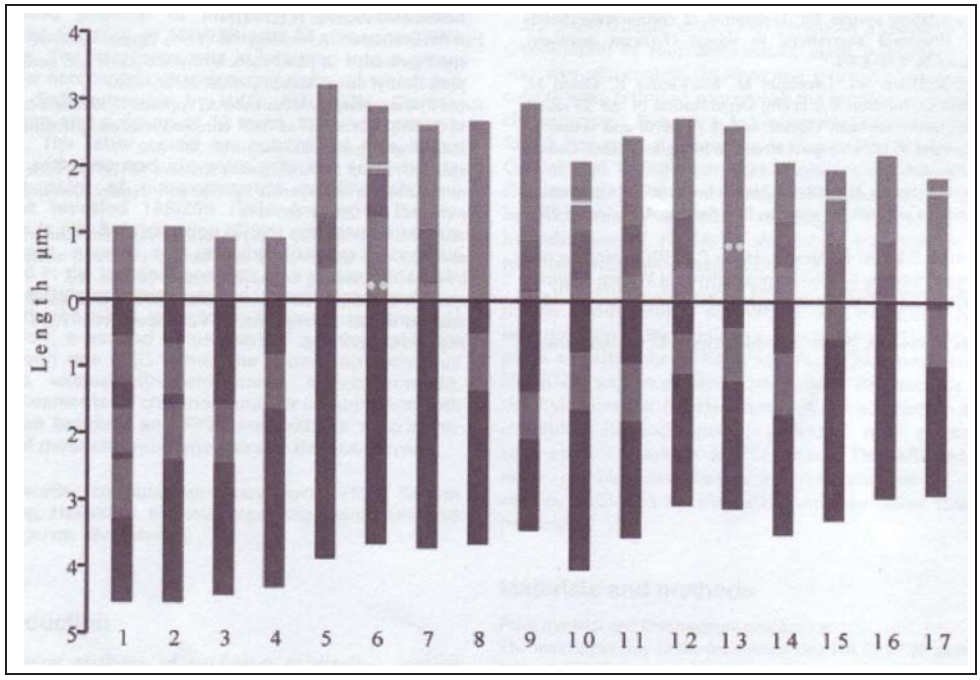


FIG. 1.8 – Idiogramme des chromosomes haploïdes de *Helianthus annuus* dérivés de la métaphase d'après Schrader *et al.* (1997).

processus physiologiques comme le remplissage du grain ou la tolérance aux maladies nous intéressent, il n'est pas indispensable de posséder l'ensemble de la séquence du tournesol, mais uniquement les régions du génome qui nous intéressent. La connaissance de l'organisation du génome du tournesol est donc indispensable.

1.2.2 Evolution des Génomes et des Gènes

L'organisation du génome du tournesol est totalement inconnue à l'heure actuelle. Très peu d'informations sont disponibles à ce sujet. On suppose que son génome est le résultat d'un événement d'alloploïdisation et qu'il est aussi en partie dupliqué. La seule certitude concernant l'organisation du génome du tournesol est qu'il résulte d'un long processus d'évolution de ces chromosomes (macro-structure) et de ces gènes (micro-structure).

Tous les organismes existant actuellement sont le résultat de l'évolution d'un organisme ancestral. Ces organismes ancestraux ont évolué de différentes façons en s'adaptant et/ou en donnant naissance à de nouvelles espèces qui elles-mêmes ont évolué en s'adaptant et/ou en donnant naissance à de nouvelles espèces et ainsi de suite. L'apparition d'une nouvelle espèce ne résulte pas d'un opération magique mais d'un très long processus évolutif du génome qui peut avoir lieu à différents niveaux, la macro-structure c'est-à-dire au niveau des chromosomes ou la micro-structure c'est-à-dire au niveau du contenu et de l'ordre des gènes, et de différentes façons, divers mécanismes : crossing-over, duplications, insertions, délétions, recombinaisons.

Evolution de la structure des génomes

L'organisation d'un génome est la conséquence de l'évolution de ce génome. Plusieurs mécanismes interviennent dans l'évolution de la macro-structure des génomes de plantes. Les deux principaux sont le croisement (ou hybridation) entre espèces (d'une même famille ou non) avec doublement des chromosomes (polyploïdisation) et les éléments mobiles (transposons et rétrotransposons).

Il existe différents type de polyploïdisation, l'alloploïdisation correspond à la formation d'une espèce issue de l'hybridation de parents d'espèces différentes, l'autopolyploïdisation correspond à la formation d'une espèce issue de l'hybridation de parents de la même espèce et les allopolyploïdes segmentées correspondent à la formation d'espèces issues de l'hybridation de parents ayant quelques remaniements chromosomiques (Leitch et Bennett, 1997). Les espèces allopolyploïdes possèdent des chromosomes dits homéologues alors que les espèces autopolyploïdes possèdent des chromosomes dits homologues. Les espèces allopolyploïdes segmentées possèdent un mélange de chromosomes homologues et homéologues. Les plantes polyploïdes ont trois ou plus de trois jeux complets de chromosomes dans leur noyau plutôt que les deux copies habituellement trouvées chez les diploïdes (Figure 1.9).

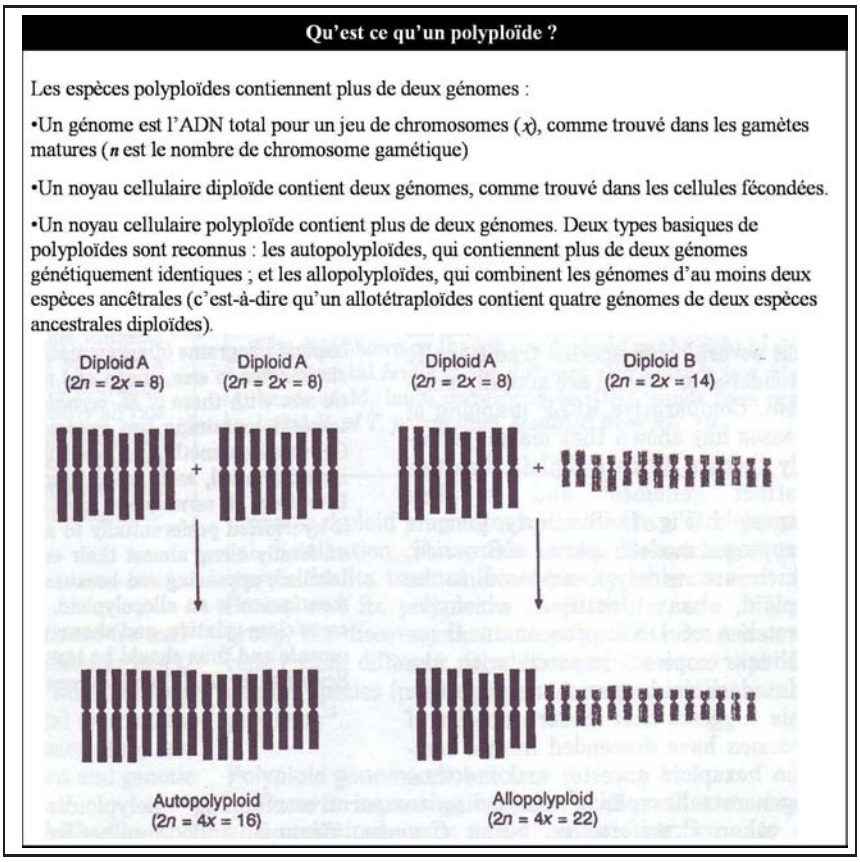


FIG. 1.9 – Qu'est-ce qu'un polyploïde, d'après Leitch et Bennett (1997).

Le colza (*Brassica napus*, par exemple, est une plante allopolyploïde qui résulte du croisement des espèces *Brassica oleracea* et *Brassica rapa* avec un doublement des chromosomes. *Brassica oleracea* et *Brassica rapa* possèdent tous 2 un génome diploïde respectivement noté AA et BB, *Brassica napus* lui possède un génome dit amphidiploïde (c'est-à-dire tetraploïde, issu d'un événement d'allopolyploïdisation) noté $AACC$.

Le blé, *Triticum aestivum*, est une plante autopolyploïde qui résulte du croisement de 3 génomes ancestraux de blé. Le blé est une plante hexaploïde qui est le résultat d'un croisement entre 3 génomes ancestraux noté AA, BB et DD avec un doublement des chromosomes, le blé a un génome noté $AABBDD$.

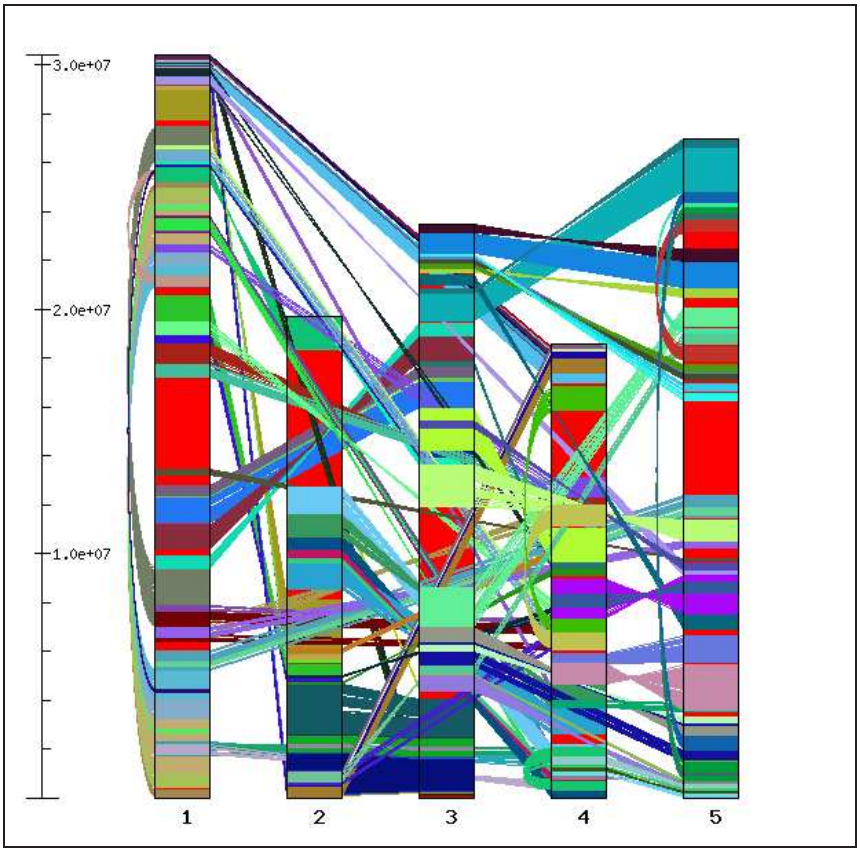


FIG. 1.10 – Représentation des différents fragments dupliqués du génome d'*Arabidopsis thaliana*, source le site web TIGR.

Des analyses faites sur *Arabidopsis thaliana*, une petite plante crucifère qui possède un génome diploïde (125 Mbases et 5 paires de chromosomes), en comparant les séquences des différents chromosomes de cette plante à l'aide de programme de dot-plot et de blast ont montrées que le génome d'*Arabidopsis* était constitué d'un patchwork de régions dupliquées (Figure 1.10, Blanc *et al.*, 2000, Vision *et al.*, 2000). Ces nombreuses duplications du génome semblent être le résultat d'au moins 2 voir 3 événements de tétraploïdisation (Simillion *et al.*, 2002, Vandepoele *et al.*, 2002, Blanc *et al.*, 2003, Ermolaeva *et al.*, 2003, Raes *et al.*, 2003). De même, pour certaines autres espèces de *Brassica* des traces d'événements de po-

lyploïdisation ont pu être observées (Song *et al.*, 1995, Johnston *et al.*, 2005). Le riz et le maïs, plantes monocotylédones, présentent aussi des traces d'événement de polyploïdisation dans leur génome (Gaut et Doebley, 1997, Ge *et al.*, 1999, Vandepoele *et al.*, 2002, Guyot et Keller, 2004, Paterson *et al.*, 2004, Zhang *et al.*, 2005). L'organisation du génome du tournesol (*Helianthus annuus* L) semble résulter d'au moins un événement d'alloploïdisation entre deux espèces différentes ($n = 8$ et $n = 9$) (Heiser et Smith, 1965). En revanche, on ne sait pas si les espèces parentales étaient diploïdes ou polyploïdes.

De nombreuses plantes présentent des traces d'événement de polyploïdisation dans leur génome. La polyploïdisation s'est produite chez au moins 70% des plantes angiospermes (Masterson, 1994), et chez 95% des ptéridophytes (Grant *et al.*, 1981). Des plantes que l'on considéraient diploïdes comme le maïs, *Arabidopsis*, ou des espèces de *Brassica* sont en fait des espèces polyploïdes ou paléopolyploïdes, ayant évolué à partir d'un ancêtre polyploïde (Leitch et Bennett, 1997). Plusieurs événements de polyploïdisation ont eu lieu au cours de l'évolution des génomes des plantes. Ainsi, *Arabidopsis thaliana* semble avoir subi au moins trois événement de polyploïdisation de son génome (Figure 1.11, Adams et Wendel, 2005). L'événement de plus ancien s'étant produit avant la divergence entre plantes monocotylédones et dicotylédones (il y a environ 225 à 300 millions d'années), ensuite avant la séparation des plantes dicotylédones (il y a environ 150 à 170 millions d'années) et enfin avant la séparation entre les genres Arabidopsis et Brassica (il y a environ 25 à 40 millions d'années).

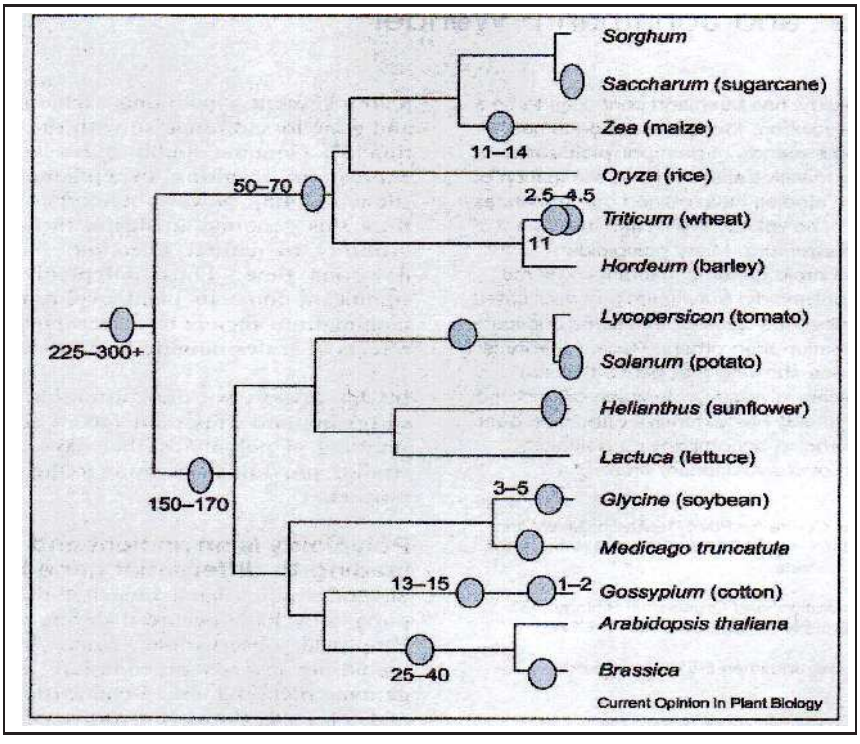


FIG. 1.11 – Représentation des événements présumés de polyploïdisation chez les angiospermes, d'après Adams et Wendel (2005).

Ces événements de polyploïdisation aboutissent généralement à la formation d'hybrides qui sont non viables ou stériles. La perte de la fertilité peut être due à une incompatibilité ou à un problème de réassociation par paire des chromosomes non spécifique lors de la méïose. Des remaniements ont alors lieu au sein du génome de ces nouvelles espèces après la polyploïdisation aboutissant parfois à la restauration de la fertilité (perte d'une région chromosomique provoquant l'incompatibilité ou duplication d'un chromosome). Ces remaniements peuvent être locaux (translocations, insertions, délétions de régions chromosomiques) ou globaux avec des évènements de diploïdisation ou d'aneuploïdisation (Wendel, 2000, Rieseberg, 2001, Wolfe, 2001, Mitchell-Olds et Clauss, 2002). Ces phénomènes ont été observés chez le maïs, le riz ou *Arabidopsis* (Gaut *et al.*, 2000, Lai *et al.*, 2004, Pontes *et al.*, 2004, Paterson *et al.*, 2004, Wang *et al.*, 2005). Parfois, ces remaniements entraînent aussi la perte de fragments comme chez le riz (Wang *et al.*, 2005) ou le blé (Chantret *et al.*, 2005).

Les éléments mobiles sont la deuxième principale source d'évolution du génome des plantes (et des animaux). Il existe plusieurs types d'éléments mobiles qui sont répartis selon deux classes (classe I et II) en fonction de leur mode de transposition. Les éléments de classe II comprennent les transposons de type *Ac/Ds*, *En/Spm*, *Mu* ou MITE qui se déplacent via une copie de l'ADN. Les éléments de classe I sont des rétrotransposons qui se déplacent en utilisant une copie ARN via une réverse transcriptase (Schmidt, 1999). Ces rétrotransposons sont divisés en deux groupes les rétrotransposons LTR (Long Terminal Repeat) et non-LTR (LINE, Long Interspersed Nuclear Elements, et SINE, Short Interspersed Nuclear Elements). Les rétrotransposons de type LTR sont principalement des rétrotransposons Ty1-*copia*-like ou Ty3-*gypsy*-like.

Ces rétrotransposons sont présents dans tout le règne végétal (Voytas *et al.*, 1992, Brandes *et al.*, 1997, Suoniemi *et al.*, 1998, Kumekawa *et al.*, 1999, Stuart-Rogers et Flavell, 2001). Ils sont présents en grand nombre dans les génomes de plantes et sont localisés sur l'ensemble du génome, dans les centromères (Heslop-Harrison *et al.*, 1999, Arabidopsis Genome Initiative, 2000, Feng *et al.*, 2002), dans les régions intergéniques (White *et al.*, 1994) et même dans d'autres transposons (SanMiguel *et al.*, 1996).

L'activité des transoposons a un impact direct sur l'organisation des génomes de plantes. La différence de taille entre les génomes de plantes est souvent due à la différence du nombre de copies de ces rétrotransposons dans les génomes. Alors que dans les petits génomes, comme *Arabidopsis*, les rétrotransposons représentent environ 4 à 5% du génome, dans les grands génomes comme le maïs ou le blé, ils peuvent représenter entre 50 et 90% du génome (Kumar et Bennetzen, 1999). Chez les plantes monocotylédones, l'un des mécanismes qui participe à l'expansion des génomes est l'insertion de rétroéléments dans les régions intergéniques (Figure 1.12). Plusieurs vagues d'invasions de rétroéléments, Figure 1.12 A, peuvent induire l'insertion de rétrotransposons au sein d'autres transposons ("nested retroelements"). Des duplications et insersions locales, Figure 1.12 B, qui ne présentent pas de rétroéléments peuvent aussi participer à l'augmentation de taille des génomes (Feuillet *et al.*, 2001).

D'un autre côté, la présence de nombreuses régions répétées dans le génome favorise aussi les crossing over inégaux, les recombinaisons homologues ou illégitimes (Figure 1.13). Ces

mécanismes entrainent ainsi la perte de fragments et la diminution de taille du génome (Devos *et al.*, 2002, Ma *et al.*, 2004, Bennetzen *et al.*, 2005). L'augmentation de l'activité des rétrotransposons est souvent observée lors d'un stress que subit la plante (réaction face à un pathogène, stress abiotique, polyploïdisation du génome).

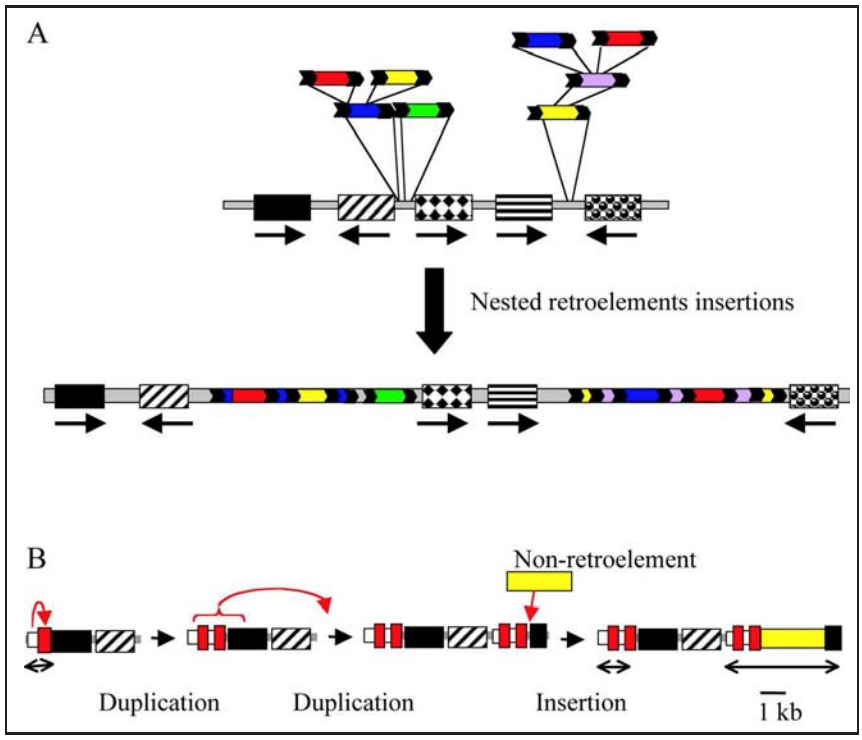


FIG. 1.12 – Mécanisme d'expansion des génomes chez les monocotylédones, d'après Feuillet et Keller (2002).

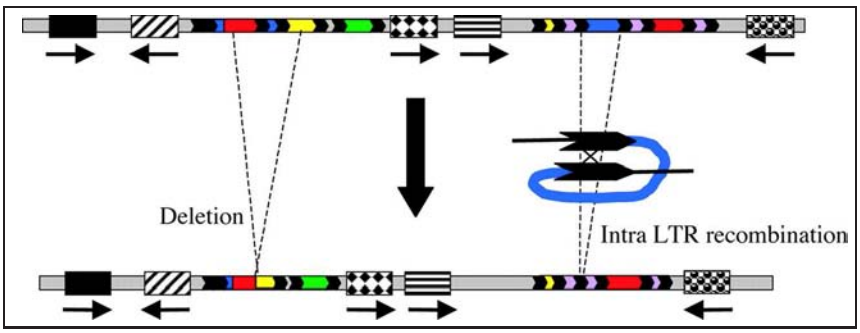


FIG. 1.13 – Mécanisme menant à la contraction des génomes, d'après Feuillet et Keller (2002).

Les transposons sont en partie responsable de la plasticité des génomes végétaux et favorisent l'insertion de gènes (souvent des ARNm) par l'intermédiaire de la réverse transcriptase au sein du génome. De nombreux remaniements, translocation et recombinaisons sont provoqués par les rétrotransposons. Le Tournesol n'échappe pas à la règle et Santini *et al.* (2002) ont montré la présence de rétrotransposons de type Ty1-*copia*-like et Ty3-*gypsy*-like dispersés le long de tous les chromosomes avec une répartition préférentielle des *gypsy*-like

dans les régions centromériques et des *copia*-like aux extrémités des chromosomes, télomères (Figure 1.14). Cependant, aucune étude n'a encore été faite sur l'impact de ces transposons sur l'organisation du génome du tournesol.

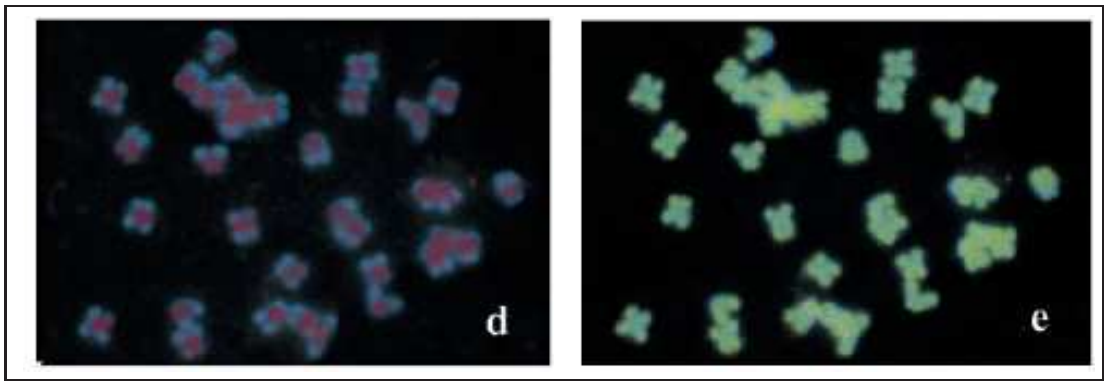


FIG. 1.14 – Hybridation au stade métaphase des chromosomes du tournesol avec les sondes des rétrotransposons de type *gypsy*-like à gauche et de type *copia*-like à droite, d'après Santini *et al.* (2002).

Un autre processus capable de modifier l'organisation du génome des plantes (mais dans une moindre mesure que la polyploïdisation ou les éléments mobiles) peut agir chez le tournesol. Le génome du tournesol peut aussi être modifié par hybridation introgressive. L'hybridation introgressive correspond à la stabilisation et au transfert de matériel génétique entre les hybrides par un back-cross avec l'un des parents (Arnold, 2004). L'hybridation introgressive permet l'adaptation de plantes à de nouveaux habitats. Les hybrides fertiles issus de croisements naturels sont fécondés par l'une des espèces parentes. L'espèce finit par récupérer les parties de génome qui lui permettent de s'adapter. L'exemple de l'espèce d'*Helianthus annuus* du Texas appelé spp. *texanus* illustre bien ce procédé. Cette plante est issue du croisement naturel entre l'espèce *Helianthus annuus* et *Helianthus debilis* spp. *cucumerifolius*. Cet hybride possède le génome de *Helianthus annuus* ainsi que trois petites régions chromosomiques qui appartiennent à *Helianthus debilis* (Kim et Rieseberg, 1999). Ce mode d'hybridation permet l'adaptation d'espèces à de nouvelles niches écologiques comme l'espèce *Helianthus anomalus* adaptée au sable des dunes et qui possède un génome correspondant à la mosaïque des génomes parentaux (Ungerer *et al.*, 1998, Schwarzbach *et al.*, 2001).

La polyploïdisation a pour conséquence la duplication de l'ensemble des gènes et les remaniements entraînent des modifications de la micro-structure. La micro-structure correspond aux différents gènes (ainsi que de l'ordre et de l'orientation de ces gènes) contenus dans un bloc chromosomique. Les gènes dupliqués peuvent être perdus localement ou par bloc, être transposés dans une autre région du génome pendant les réarrangements de celui-ci (Lai *et al.*, 2004) modifiant ainsi la micro-structure du génome. Blanc et Wolfe (2004) ont montré que, chez *Arabidopsis*, les gènes n'étaient pas perdus au hasard mais en fonction de la catégorie fonctionnelle à laquelle ils appartiennent. Les gènes dupliqués impliqués dans les voies de

transduction et de transcription de signaux sont préférentiellement maintenus, alors que les gènes dupliqués impliqués dans la réparation de l'ADN sont préférentiellement perdus. Les mêmes résultats ont été observés chez la levure par Seoighe et Wolfe (1999, 1999). D'autres mécanismes influencent aussi la micro-structure du génome des plantes. Les crossing-over inégaux favorisent l'augmentation ou la diminution des gènes dupliqués regroupés en cluster. Ces gènes en cluster subissent une pression de sélection avec des rapides taux de "naissance" et de "mortalité" de ces nouveaux gènes dupliqués. Les éléments mobiles peuvent aussi favoriser la duplication en tandem des gènes par leur insertion et leur excision.

Les génomes évoluent donc par duplication (complète ou partielle des chromosomes), translocation, inversion, délétion de chromosomes ou fragments chromosomiques. Tous ces remaniements ont pour conséquence la formation de génome dont l'organisation est un patchwork des blocs chromosomiques délétés, dupliqués, inversés, transloqués d'un ou de plusieurs génomes ancestraux. Tous ces mécanismes participent à l'évolution du génome au niveau chromosomique (macro-structure) mais aussi au niveau de la présence, de l'ordre et du sens des gènes contenus dans ces blocs (micro-structure).

Evolution des gènes

Les gènes dupliqués subissent des modifications dans leur fonction et/ou leur expression ainsi que dans leur structure. Lors de la duplication des gènes (polyploïdisation, transposons, etc.), la fonction et la structure des gènes dupliqués ne sont pas modifiées. Par contre, l'expression de ces gènes peut subir des modifications. Ainsi, des effets épigénétiques peuvent se mettre en place et entraîner le silencing ou la régulation modifiée de l'expression des gènes dupliqués (Comai *et al.*, 2000, Adams *et al.*, 2003, Wang *et al.*, 2004, Adams *et al.*, 2004). Les gènes dupliqués exprimés peuvent aussi être spécifiques d'un organe ou s'exprimer dans des conditions particulières, ainsi Blanc et Wolfe (2004) ont montré que chez *Arabidopsis* parmi des gènes dupliqués 57% des gènes récemment dupliqués et 73% des gènes anciennement dupliqués avaient des régulations divergentes. De plus, Blanc et Wolfe (2004) ont aussi montré que les gènes d'*Arabidopsis* dupliqués d'un réseau de régulation avaient évolué parallèlement formant ainsi deux réseaux parallèles chacun contenant un gène de chaque paire dupliquée avec une forte corrélation dans l'expression des gènes d'un même réseau et une faible corrélation dans l'expression des gènes dupliqués. La duplication de ces gènes a entraîné la mise en place de deux réseaux similaires mais régulés de façon différente.

Les nombreuses modifications (duplications, réarrangements) successives subies par les gènes au cours de l'évolution du génome entraînent à terme la formation de familles multigéniques. La stabilisation du génome après la duplication entraîne, à plus long terme, l'évolution des gènes au niveau de l'ADN. Les gènes dupliqués semblent évoluer indépendemment de leur gène ancestral (Cronn *et al.*, 1999). Ces modifications de la séquence ADN permettent aux gènes dupliqués d'évoluer et de différer dans leur fonction par néofonctionnalisation (acquisition d'une nouvelle fonction) ou par sous-fonctionnalisation (récupération

d'une des fonctions du gène ancestral, Force *et al.*, 1999). Les familles multigéniques finissent par avoir des rôles fonctionnels différents (Cannon *et al.*, 2004).

Toutes ces duplications et ces remaniements rendent complexe la relation existant entre les gènes. Afin de pouvoir distinguer les gènes au sein d'un organisme et entre organismes, une terminologie a été mise en place (Fitch, 2000). Les gènes sont dits homologues quand ils descendent d'un même gène ancestral (avec quelques divergences dû à l'évolution de la séquence nucléotidique) et sont opposés aux gènes dits analogues, gènes qui sont similaires mais ne descendant pas d'un même gène ancestral. La duplication et la divergence des séquences des séquences est l'un des mécanismes de l'évolution des gènes, qui, avec la spéciation, entraînent la formation de séquences homologues (Figure 1.15). Parmi les gènes homologues, on distingue les gènes dits orthologues et paralogues. Des gènes dits orthologues sont des gènes similaires appartenant à des espèces différentes mais descendant d'un ancêtre commun, c'est-à-dire que ce sont des gènes qui dérivent d'un évènement de spéciation et non de duplication alors que des gènes dits paralogues sont des gènes issus d'un évènement de duplication au sein du génome d'une espèce. Ces gènes évoluent de façons différentes mais de trop grandes modifications de leur séquence entraîneraient une perte de fonction. C'est pourquoi, la structure de ces gènes (découpage intron/exon) est fortement conservée entre gènes orthologues ou paralogues bien que la séquence évolue différemment (mutations, insertions/délétions). La conservation des positions des introns au sein des séquences génomiques a été étudiés pour plusieurs gènes ou familles de gènes paralogues ou orthologues (Sahrawy *et al.*, 1996, Proudhon *et al.*, 1996, Lunn, 2003). Carels et Bernardi (2000) ont montré que la position des introns était généralement conservée entre gènes paralogues et orthologues chez les plantes mais par contre la taille de ces introns peut être très différentes.

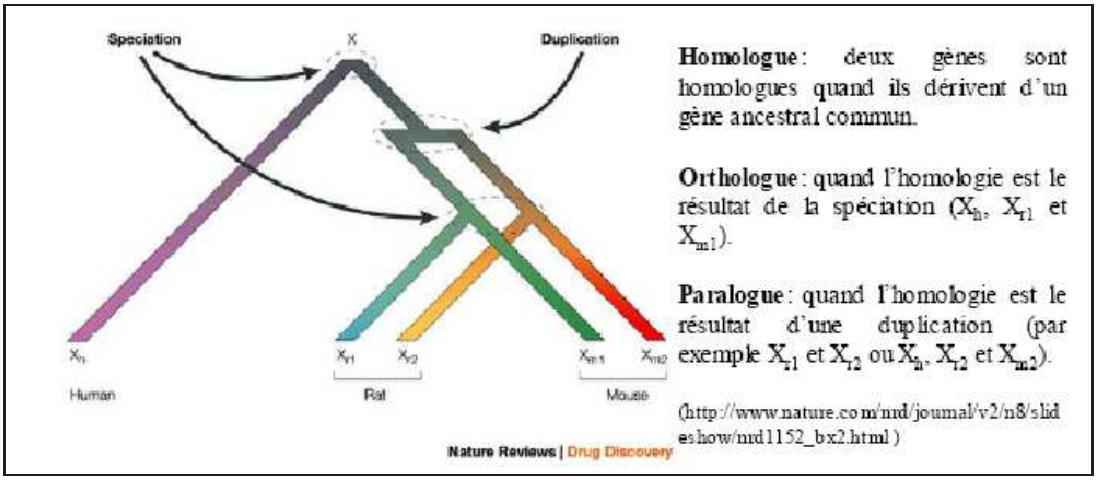


FIG. 1.15 – Evolution des relations entre gènes et terminologie.

Conclusion sur l'évolution des génomes et des gènes

La formation d'espèces hybrides, les remaniements chromosomiques et les éléments mobiles participent tous à la plasticité des génomes végétaux. L'organisation des génomes de plantes est donc continuellement en évolution. Mais cette évolution est lente car elle se produit sur des millions d'années. A l'échelle du temps, ces évolutions sont fréquentes, mais à l'échelle humaine, l'organisation du génome des plantes est plutôt figée. D'un côté, c'est un désavantage, car l'étude de l'évolution des génomes relève plus de la paléogénétique ou paléogénomique que du suivi des génomes de plantes issues de croisements tests. D'un autre côté, cette relative stabilité des génomes permet d'étudier la position des différents blocs chromosomiques qui ont évolué entre espèces proches ou entre espèces éloignées.

Les études évolutives sont très intéressantes car elles permettent de débloquer certaines situations. Une étude trop centrée sur l'analyse d'un ou plusieurs gènes chez un seul organisme peut vite devenir ardue, mais cette même étude faite avec plusieurs organismes différents peut permettre de reconstituer plus facilement l'historique évolutif de ces gènes et de faciliter l'interprétation de ces analyses, surtout lorsqu'il y a des pertes de gènes. L'inconvénient majeur de ce genre d'étude évolutive est qu'elle nécessite de posséder les informations pour plusieurs organismes dont au moins un ayant fait l'objet d'une étude poussée du gène afin de faciliter la transposition à d'autres séquences. C'est pourquoi de nombreux projets de séquençage (du génome complet, des ARNm, etc.) d'organismes ont été entrepris afin de transposer les informations d'un organisme à un autre. Ces dix dernières années, il y a eu un réel essor des études de comparaison de génomes qui ont permis de mieux comprendre l'organisation et l'évolution des génomes de plantes.

1.2.3 Comparaison de Génomes et Synténie

Tous les organismes descendent d'ancêtres communs. L'évolution de leur génome a entraîné de nombreuses divergences qui sont à la base de la diversité de formes, de fonctions et d'adaptation des organismes vivants. Ces divergences sont d'autant plus importantes que les espèces sont éloignées. Grâces aux progrès de la biologie moléculaire et de la génétique, il est aujourd'hui possible de comparer les génomes entiers de différents organismes, ce qui a mené à l'essor d'un nouveau domaine au sein de la génomique : la comparaison de génomes.

Les études de comparaison de génomes sont très récentes et s'appuie sur 2 conditions : que le contenu et l'ordre des gènes dans le génome des espèces actuelles dérivent de celui d'un ancêtre commun et que les informations sur les espèces actuelles servent à rétablir l'ordre et le contenu des gènes du génome ancestral. La comparaison de génomes peut être faite à 2 échelles, selon que l'on s'intéresse aux cartes génétiques ou aux cartes physiques.

Dans le premier cas, la comparaison de cartes génétiques s'appuie sur la comparaison de la position et de l'ordre de marqueurs moléculaires communs entre 2 ou plusieurs espèces et

permet ainsi d'avoir une vision globale sur l'ensemble du génome des réarrangements et des conservations de blocs synténiques au sein des génomes comparés. Ainsi, les cartes génétiques de 2 ou plusieurs espèces phylogénétiquement proches sont alignées à l'aide de marqueurs orthologues et l'unité de distance génétique utilisée est le centiMorgan, cM (pourcentage de recombinaison). La Figure 1.16 présente la cartographie génétique comparée du chromosome 2 d'*Arabidopsis thaliana* avec les groupes de liaisons 3 et 6 de la betterave à sucre. Les marqueurs orthologues (loci qui dérivent d'un même locus andestral par spéciation) nous permettent de comparer le génome de différentes espèces (Barnes, 2002).

Dans le deuxième cas, la comparaison de cartes physiques s'appuie sur la comparaison des séquences d'ADN de 2 ou plusieurs espèces, elle peut être faite sur l'ensemble du génome lorsque ces organismes ont été entièrement séquencés ou sur une partie du génome (clones BAC, par exemple) et permet d'estimer la conservation de synténie au niveau des gènes pour l'ensemble des gènes présents sur le fragment de la séquence d'ADN. Ainsi, les cartes physiques sont comparées en utilisant les séquences d'ADN de 2 ou plusieurs espèces et la distance est mesurée en paires de bases d'ADN (distance réelle). La Figure 1.16 présente la cartographie physique comparée des fragments d'ADN de 2 espèces A et B. Il s'agit d'une région homologue qui dérive du même fragment ancestral. Dans l'espèce B ce fragment a été dupliqué au cours de l'évolution. Par rapport à la région ancestrale des délétions, insertions, inversions sont observées chez les espèces actuelles (Barnes, 2002).

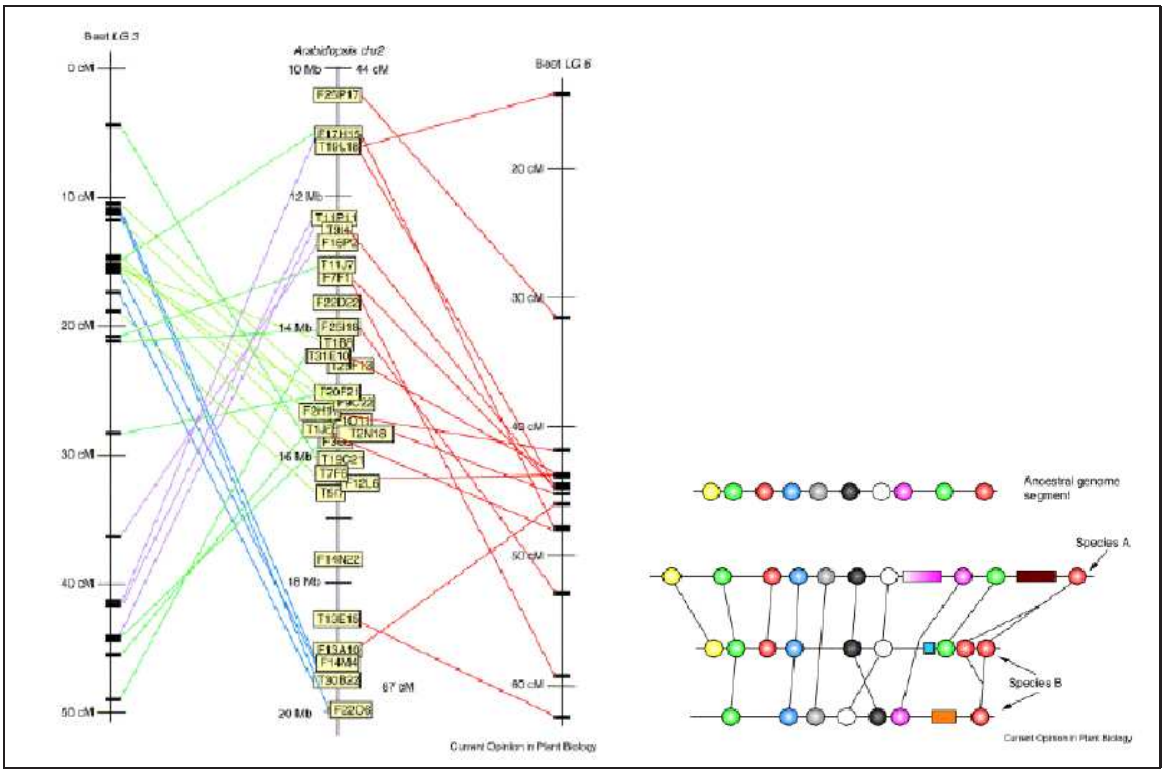


FIG. 1.16 – Cartographie génétique comparée (à gauche) et cartographie physique comparée (à droite), Barnes, 2002.

Les premières études de comparaison de génomes ont été faites chez les animaux en comparant des marqueurs moléculaires de cartes génétiques, de cartes d'hybrides de radiation et par hybridation sur chromosome (chromosome painting) entre le génome humain et des espèces proches toutes étant des mammifères comme les rongeurs (Nadeau et Taylor, 1984, Cheng *et al.*, 1988, Watkins-Chow *et al.*, 1997, O'Brien *et al.*, 1999), les ruminants (Echard *et al.*, 1994, Gautier *et al.*, 2002, Schibler *et al.*, 1998) ou le porc (Robic *et al.*, 1999, 2001, Rink *et al.*, 2002). Ces études ont révélé des conservations de blocs de marqueurs (macrosynténie) entre des espèces animales (Chowdhary *et al.*, 1998).

Chez les plantes, les cartes génétiques d'un grand nombre d'espèces végétales ont été construites, ce qui a permis le développement de la cartographie génétique comparée à partir de la fin des années 1980 (Schmidt, 2000).

Les premières expériences de comparaison de cartes génétiques ont été faites sur des espèces de la famille des *Solanaceae* (*Lycopersicon esculentum*, *Solanum tuberosum*) en utilisant des marqueurs RFLP dérivés de la tomate (Bonierbale *et al.*, 1988, Gebhardt *et al.*, 1991, Tanksley *et al.*, 1992, Livingstone *et al.*, 1999). La comparaison des cartes génétiques de la tomate et la pomme de terre a mis en évidence l'homologie des 12 groupes de liaison, la conservation de l'ordre des marqueurs et une différence remarquable au niveau de la fréquence de recombinaison (1276 cM chez la tomate et 684 cM chez la pomme de terre). Cinq inversions paracentriques qui différencient les 2 espèces ont été mises en évidence (Bonierbale *et al.*, 1988, Tanksley *et al.*, 1992). De même, la comparaison entre les cartes génétiques du poivron, de la tomate, de la pomme de terre et de l'aubergine a permis d'identifier quelques mécanismes d'évolution du génome de cette famille (Livingstone *et al.*, 1999, Doganlar *et al.*, 2002). Une trentaine de cassures, comprenant des translocations, des inversions paracentriques et péricentriques, seraient suffisants pour expliquer les différences entre les 4 génomes. Malgré la taille physique plus élevée chez le poivron et l'aubergine, expliquée par la présence de rétro-tranposons chez le poivron, le contenu en gènes est conservé. Ces études de comparaison de cartes génétiques ont montré qu'il existait une conservation de la macrosynténie entre les différentes espèces de *Solonaceae*, même si quelques remaniements ont lieu au sein des génomes.

Depuis, d'autres études ont été faites sur d'autres familles notament chez les *Brassicaceae* dont fait partie la plante modèle *Arabidopsis thaliana*. Les espèces *Brassica* sont diploïdes et dériveraient d'un ancêtre commun ayant 6 chromosomes (*Brassica rapa*, génome AA ; *Brassica nigra*, génome BB ; *Brassica oleracea*, génome CC). Par hybridation ces espèces sont à l'origine des espèces amphidiploïdes (*Brassica napus*, génome AACC ; *Brassica carinata*, génome BBCC ; *Brassica juncea*, génome AABB). La comparaison des 3 génomes diploïdes (*B. nigra*, *B. oleracea* et *B. rapa*) a permis de mettre en évidence la nature dupliquée de ces génomes (trois copies d'un génome ancestral sont identifiables), et un minimum de 24 réarrangements entre les 3 génomes se seraient produits (Lagercrantz et Lydiate, 1996). La comparaison entre *Arabidopsis thaliana* et d'autres espèces de *Brassica* a montré que la macro et microcolinéraité sont très conservées avec *Capsella rubella* (Acarkan *et al.*, 2000). Des fragments de plus de 60 cM sont entièrement conservés. De même en comparant les cartes génétiques d'*Arabidopsis thaliana* et de *Brassica oleracea*, 26 réarrangements ont été mis en évidence et la longueur des

fragments colinéaires varie entre 4 et 50 cM (Kowalski *et al.*, 1994). Environ 90 réarrangements ont été révélés entre *Arabidopsis thaliana* et *Brassica nigra* avec des fragments conservés qui ont une longueur moyenne de 8 cM (Lagercrantz, 1998, Figure 1.17).

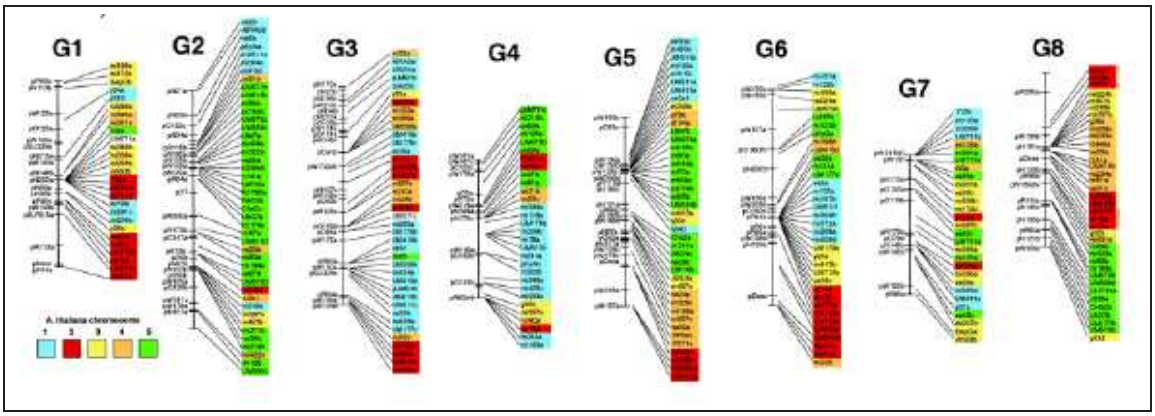


FIG. 1.17 – Comparaison entre la carte génétique d'*Arabidopsis thaliana* et celle de *Brassica nigra*, Lagercrantz, 1998.

Il en va de même avec les espèces du genre *Poaceae* (Kilian *et al.*, 1995, Bennetzen et Freeling, 1997). La comparaison de cartes génétiques chez le blé a permis de démontrer la colinéarité qui existe entre les 3 génomes A, B et D du blé hexaploïde (Chao *et al.*, 1989). Hulbert *et al.* (1990) ont ensuite démontré que les sondes développées chez le maïs pouvaient s'hybrider avec le sorgho, le millet et la canne à sucre. Ils ont ainsi produit le premier alignement des groupes de liaison entre maïs et sorgho qui montrait la similarité des 2 génomes. Ahn et Tanksley (1993) ont pu aligner les cartes génétiques du riz et du maïs démontrant que le contenu et l'ordre des marqueurs étaient identiques dans certains chromosomes ou régions et que la plupart des gènes étaient dupliqués chez le maïs. Cette conservation de l'ordre des gèns dans certaines régions chromosomiques ressort également des comparaisons entre blé et riz (Kurata *et al.*, 1994), blé, orge et seigle (Devos *et al.*, 1993), sorgho et maïs (Pereira *et al.*, 1994), sorgho et canne à sucre (Guimaraes *et al.*, 1997, Ming *et al.*, 1998), riz et sorgho (Ventelon *et al.*, 2001). A partir de la comparaison des cartes génétiques des différentes espèces de plantes monocotylédones, Gale et Devos (1998) ont pu aligner les génomes du blé, du maïs, du riz, de l'avoine, du sorgho, de la canne à sucre et du millet (Figure 1.18). Trente blocs à peine du génome du riz sont suffisants pour décrire le génome des autres espèces (Devos et Gale, 2000, Devos, 2005). Malgré la différence du nombre de chromosomes (de 5 à 12) et du contenu en ADN (de 400 à 6 000 Mbases) des espèces comparées, peu de réarrangments ont été révélés. Il faut, néanmoins, souligner que le nombre de réarrangements entre différents génomes varie beaucoup selon les lignées e qu'il ne serait pas relié à la distance évolutive entre les espèces de manière certaine (Devos et Gale, 2000).

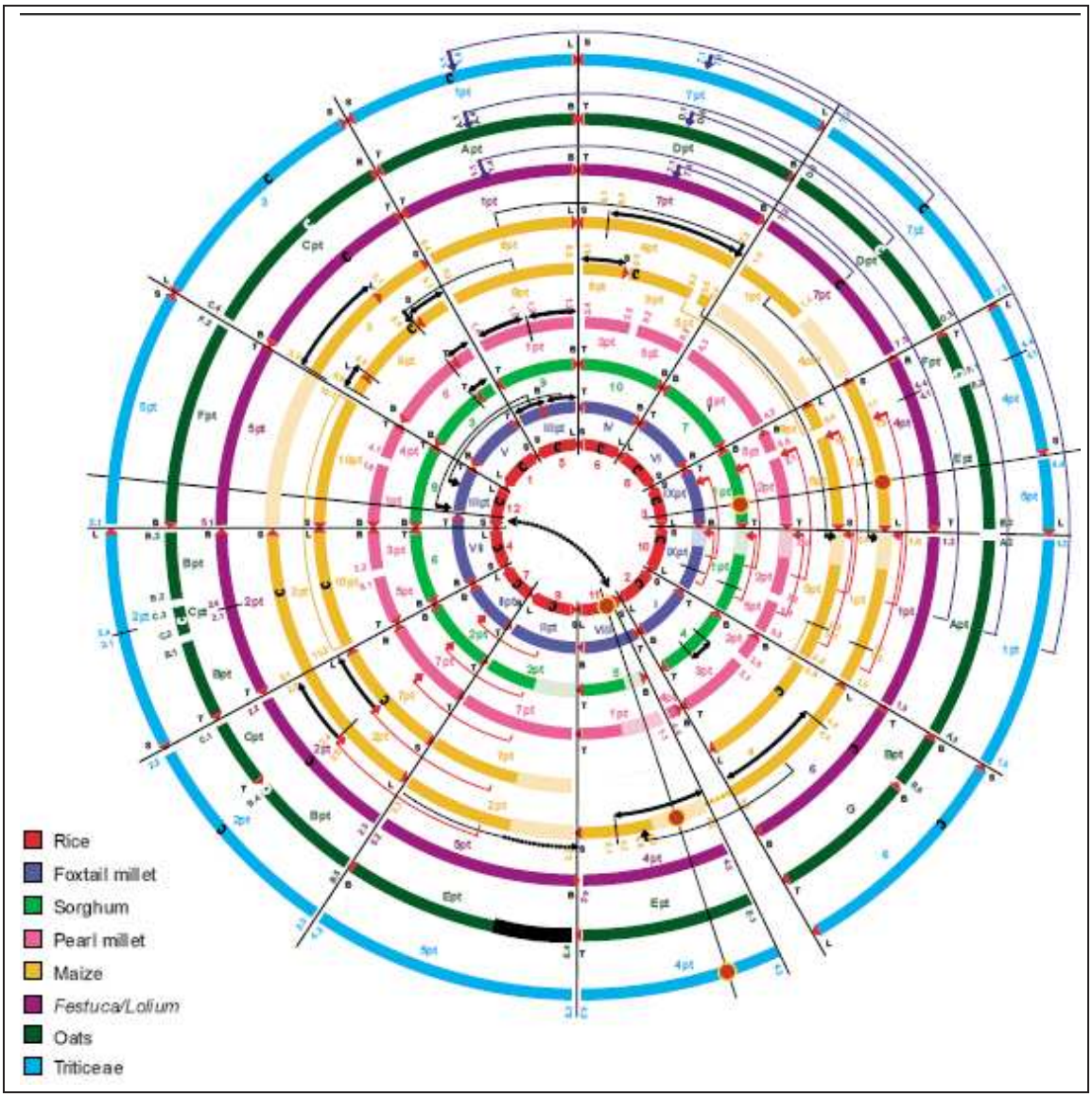


FIG. 1.18 – Diagramme du 'Crop Circle' montrant les relations existantes à l'heure actuelle entre les génomes de 8 espèces appartenant à 3 sous-familles différentes, Devos, 2005.

Ces comparaisons de cartes génétiques entre espèces animales et entre espèces végétales ont montré qu'il existait une conservation de la macrosynténie entre espèces proches avec une meilleure conservation des marqueurs entre les espèces animales qui pourtant sont aussi anciennes que les espèces végétales dicotylédones (Arnason *et al.*, 1996, Janke *et al.*, 1997). Cependant, la macrosynténie ne peut être explorée qu'entre espèces végétales proches à cause des limites techniques imposées par les marqueurs moléculaires utilisés (ce problème ne se pose pas pour l'étude de la macrosynténie entre mammifères car l'évolution des génomes chez les animaux semble différer de celle des végétaux et entraînerait moins de remaniements au sein des génomes animaux ; en revanche ce problème est observé en comparant des espèces animales plus éloignées comme homme / poulet ou homme / fugu).

La comparaison de cartes génétiques chez les végétaux se heurte à 2 difficultés principales. La première concerne la disponibilité de marqueurs moléculaires orthologues facilement transférable d'une espèce à l'autre et suffisamment nombreux pour être répartis sur tout le génome. La seconde concerne la possibilité d'identifier des segments conservés à l'aide de la cartographie génétique. Les familles multigéniques et les pseudogènes, par exemple, peuvent parfois rendre difficile l'identification du bon gène orthologue, surtout si les espèces comparés sont éloignées phylogénétiquement. En effet dans ce cas, il est difficile de distinguer l'érosion de l'identité de séquence due à la spéciation de celle due à la duplication. De plus, la cartographie génétique, il ne faut pas l'oublier, n'est que l'estimation de la fréquence de recombinaison et de l'ordre des marqueurs et qu'elle est toujours affectée par une incertitude statistique. Il est donc judicieux d'être prudent sur les conclusions qui peuvent être tirées d'une étude de cartographie comparée, en particulier si les marqueurs communs cartographiés entre 2 espèces ne sont peu nombreux. De plus, la distance entre deux marqueurs sur une carte génétique ne correspond qu'à un pourcentage de recombinaison et ne reflète pas du tout le contenu en gènes et la séquence d'ADN de cette région du génome. Entre 2 marqueurs peuvent se cacher des centaines de gènes, et la macrosynténie entre deux espèces n'implique pas forcément l'existence de conservation des gènes, de leur ordre et de leur sens (microsynténie).

Même si la cartographie génétique apporte des éclaircissements importants sur la structure, l'organisation et sur le mode et la vitesse d'évolution des génomes de plantes (Schmidt, 2000), il devient de plus en plus clair que ce n'est que la comparaison directe des séquences d'ADN de génomes entiers qui pourra élucider les processus de l'évolution de ceux-ci et le rôle joué par les réarrangements chromosomiques dans la spéciation (Mitchell-Olds et Clauss, 2002, Sankoff et Nadeau, 2003). L'étude de la microsynténie requiert les séquences génomiques et leur annotation pour comparer les différents génomes.

De nombreux programmes de séquençage de génomes ont alors été réalisés afin de permettre les comparaisons de la microsynténie entre espèces. Le séquençage du génome de la levure (Goffeau *et al.*, 1996), de la drosophile (Adams *et al.*, 2000), de *C. elegans* (The C. elegans Sequencing Consortium, 1998), du fugu (Taylor *et al.*, 2002), de l'homme (Dunham *et al.*, 1999, The Chromosome 21 mapping and sequencing consortium, 2000, International Human Genome Sequencing Consortium, 2001, Craig Venter *et al.*, 2001, Mungall *et al.*, 2003) et d'*Arabidopsis thaliana* (Lin *et al.*, 1999, Mayer *et al.*, 1999, Salanoubat *et al.*, 2000, Tabata *et al.*, 2000, Theologis *et al.*, 2000) a été achevé, mais d'autres organismes sont en cours

de séquençages (*Medicago truncatula*, le riz, le maïs, etc.) ainsi qu'un nombre croissant de programmes de génomique d'autres espèces animales et végétales (séquençages d'ARNm, de séquences EST, de fragments d'ADN génomique).

Le séquençage de génome complet et de fragments de génomes (notament de clones BAC) a permis la comparaison de cartes physiques entre différentes espèces aussi bien animales que végétales. Le nombre de séquences génomiques disponible chez les animaux est beaucoup plus important que chez les végétaux ce qui entraîne que les comparaisons de génomes chez les animaux est bien plus avancé que chez les végétaux. Les comparaisons de cartes physiques chez les animaux ont démontré la conservation de la microsynténie entre mammifères. Les études de microsynténie ont confirmé que la colinéarité est aussi généralement observée au niveau des gènes entre différentes espèces. La microsynténie chez les animaux est conservée aussi bien entre espèces mammifères, avec quelques réarrangements (Oakey *et al.*, 1992, Goureau *et al.*, 2001, Lopez-Corrales *et al.*, 1998, Martins-Wess *et al.*, 2002) qu'entre espèces mammifères et d'autres espèces plus distantes comme le poisson globe ou le poulet (Elgar *et al.*, 1996, McLysaght *et al.*, 2000, Thomas *et al.*, 2003).

Pour les plantes, la comparaison de cartes physiques a principalement été faite en comparant entre elles les plantes monocotylédones (avec le riz en plante de référence) sur des fragments de génome macrosynténique ainsi qu'entre *Arabidopsis thaliana*, plante de référence, et d'autres espèces végétales (espèces appartenant au *Brassicaceae*, le riz, *Medicago truncatula*). La microsynténie a ainsi pu être mise en évidence chez des espèces de la même famille comme chez les *Poaceae* (Feuillet et Keller, 1999, 2002, Bennetzen et Ramakrishna, 2002, Song *et al.*, 2002). La comparaison de séquences de plusieurs clones BAC orthologues entre le génome du riz, du blé (diploïde), du sorgho et de l'orge a permis de mettre en évidence la conservation de l'emplacement et de l'ordre des gènes dans cette région ainsi que la différence de taille entre les régions orthologues du riz et du sorgho (génome plus compact) et celles de l'orge et du blé (Figure 1.19).

La disponibilité de la séquence complète du génome d'*Arabidopsis* a permis de comparer de génome de cette plante aux séquences de fragments d'ADN génomiques partiels de différentes espèces végétales. La comparaison entre *Arabidopsis* et des espèces appartenant au *Brassicaceae* a permis de mettre en évidence une conservation de la microsynténie avec quelques remaniements, insertion / perte de gène ou duplication de gènes (Acarkan *et al.*, 2000, O'Neill et Bancroft, 2000, Parkin *et al.*, 2002, Li *et al.*, 2003). De même entre *arabidopsis*, la plante modèle, et des espèces *Poaceae* (van Dodeweerd *et al.*, 1999, Mayer *et al.*, 2001, Salse *et al.*, 2002) ou des espèces appartenant à d'autres familles comme avec les *Fabaceae* (Grant *et al.*, 2000, Foster-Hartnett *et al.*, 2002, Yan *et al.*, 2003), ou les *Solanaceae* (Ku *et al.*, 2000, Rossberg *et al.*, 2001, Gebhardt *et al.*, 2003). La Figure 1.20 présente la comparaison de régoins génomiques orthologues de différentes espèces végétales, *Arabidopsis*, *Capsella* et la tomate (Rossberg *et al.*, 2001). Cette comparaison montre la conservation de microsynténie au sein de cette région avec une conservation des différents gènes et de leur ordre même si quelques réarrangments peuvent être observés (inversion du sens du gène A chez la tomate, et inversions de position des gènes C et D chez la tomate par rapport à *Arabidopsis* et *Capsella*).

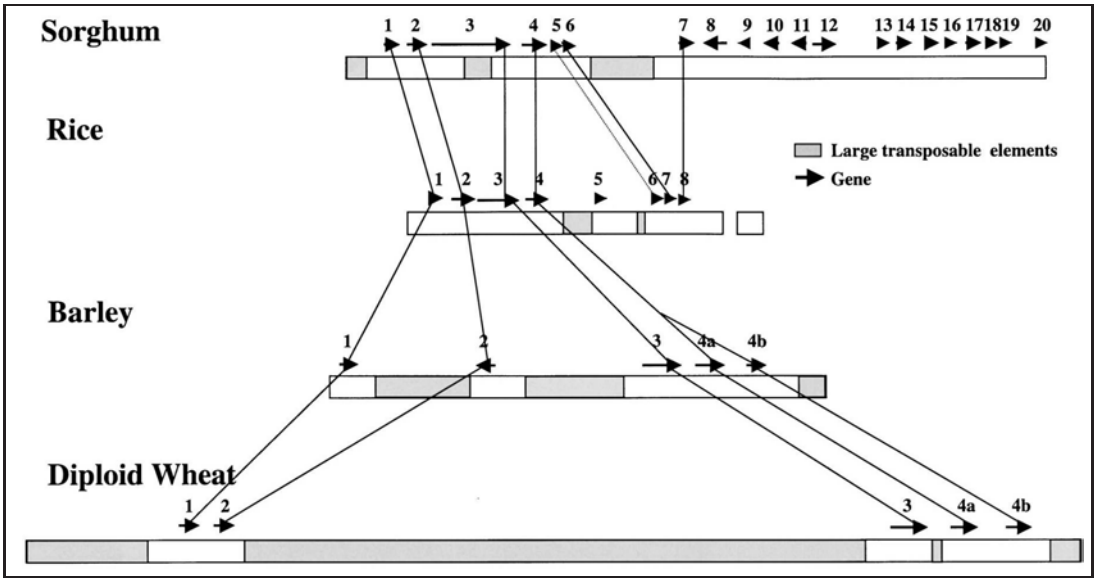


FIG. 1.19 – Comparaison d'une région orthologue d'orge, de riz, de sorgho et de blé diploïde, Ramakrishna *et al.* (2002).

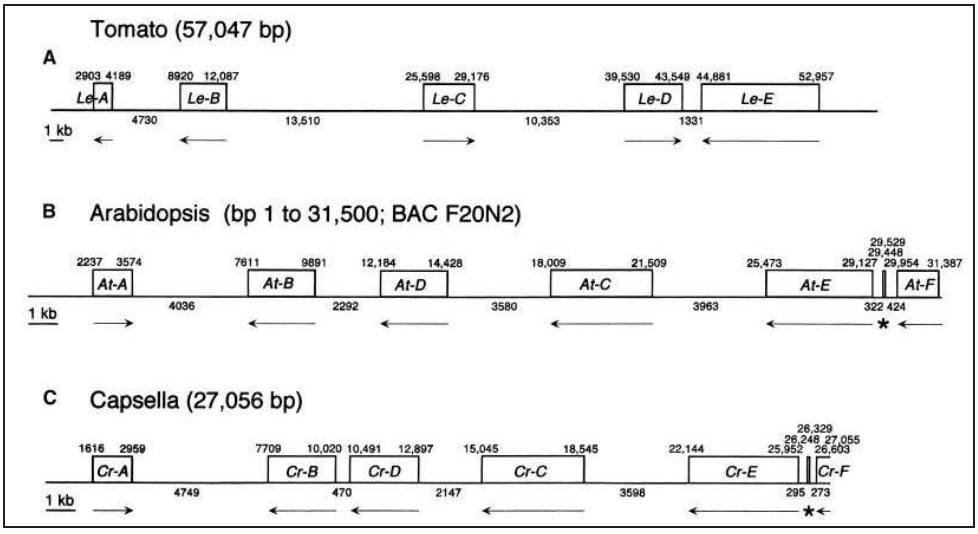


FIG. 1.20 – Comparaison de l'arrangements des gènes dans les régions génomiques orthogues d'*Arabidopsis*, de *Capsella* et de la tomate, Rossberg *et al.* (2001).

Les plantes ont évolué à partir d'un ancêtre commun polyploïde. Les relations de synténie existant aujourd'hui entre les espèces angiospermes sont le résultat des remaniements, duplications, translocations, pertes de gènes issus de l'évolution du génome de cet ancêtre commun (Bancroft, 2001, Simillion *et al.*, 2002, Ermolaeva *et al.*, 2003, Bowers *et al.*, 2003). Les différentes évolutions des génomes ancestraux ont conduit à la formation de nouvelles espèces. Ces espèces possèdent un génome dont l'organisation résulte d'un patchwork de régions chromosomiques dupliquées, inversées, délétées, transloquées – macro-structure – et chacune de ces régions a subi des modifications, pertes de gènes, duplications en tandem – micro-structure – (Blanc *et al.*, 2000). En plus de ces remaniements internes, les éléments répétés (transoposons et rétrotransposons) viennent s'insérer au cœur du génome ajoutant une difficulté supplémentaire dans l'organisation de ceux-ci.

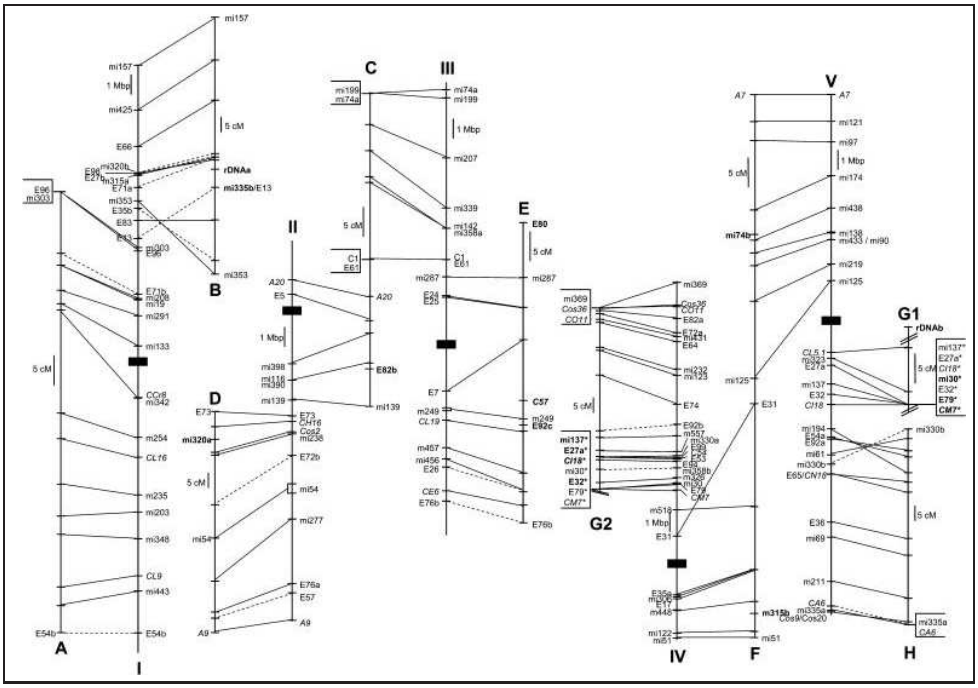


FIG. 1.21 – Comparaison entre la séquence du génome d'*Arabidopsis thaliana* et la carte génétique de *Capsella rubella*, Boivin *et al.* (2004).

Les organisations de génome peuvent être comparées entre organismes afin de déterminer si la synténie est conservée. La macro- et microsynténie sont plus ou moins fortement conservées entre espèces en fonction du temps de divergence ce qui permet le transfert d'information d'un organisme à un autre même si quelques réarrangements peuvent avoir eu lieu au sein des gènes (Schmidt, 2000, Bancroft, 2001, Barnes, 2002). Plusieurs génomes d'espèces végétales sont maintenant étudiés en utilisant les connaissances acquises chez *Arabidopsis*. A partir des informations obtenues chez la plante modèle, des gènes d'intérêt chez la tomate ont pu être localisé dans le génome de celle-ci grâce à la synténie entre le génome d'*Arabidopsis* et celui de cet organisme (Ku *et al.*, 2001, Oh *et al.*, 2002). La séquence complète du génome d'*Arabidopsis* et son annotation ont aussi permis d'énormément faciliter l'analyse des génomes

des plantes appartenant au *Brassica* (Fourmann *et al.*, 2002, Babula *et al.*, 2003, Lukens *et al.*, 2003). Des marquers moléculaires définis à partir d'*Arabidopsis* ont ainsi pu être utilisés pour construire une carte génétique de *Capsella rubella* et d'établir la macrosynténie entre ces 2 organismes (Figure 1.21, Boivin *et al.*, 2004).

Ces transferts d'informations permettent de faciliter les études de structure et d'organisation de génomes et de mieux connaître les différents gènes contenus dans les génomes végétaux. De nombreux projets de séquençages de séquences EST et ARNm ont été ou sont en cours de réalisation pour différents organismes végétaux afin de justement pouvoir profiter des connaissances acquises sur des organismes dits modèles. Le génome du tournesol (3 000 Mb) est très grand comparé à celui des plantes modèles *Arabidopsis* (125 Mb) ou du riz (450 Mb), mais il doit sûrement être synténique (ou du moins en partie) avec les génomes de ces plantes.

1.3 Objectifs du Travail de Thèse

Les analyses d'organisation de génomes de plantes comme *Arabidopsis*, le riz ou le maïs ont montré que deux organisations de génome sont envisageables chez le tournesol : les gènes sont espacés régulièrement comme chez *Arabidopsis* (avec un espace intergénique plus important) ou les gènes sont plus ou moins regroupés en îlot comme chez le maïs ou le blé avec de nombreuses régions pauvres en gènes (Bennetzen *et al.*, 1998, Conley *et al.*, 2004). Si les gènes sont espacés régulièrement tout au long du génome, alors dans ce cas là, il est nécessaire de séquencer la totalité du génome pour avoir l'ensemble des gènes. Mais si les gènes sont regroupés en cluster, il est envisageable de ne séquencer que ces régions riches en gènes. Connaître l'organisation du génome du tournesol est donc une étape préalable nécessaire afin de déterminer la meilleure stratégie de séquençage à mettre en place avant d'envisager de lancer ce type de projet chez le tournesol.

Mes objectifs de thèse sont de définir et de mettre en place une méthodologie et de pouvoir évaluer l'organisation du génome du tournesol et le niveau de synténie existant entre son génome et celui de la plante modèle *Arabidopsis thaliana*. La mise en place de la méthodologie s'appuie sur les connaissances déjà acquises en comparaison de génome et en synténie pour différents organismes. J'ai utilisé les informations de génomes bien étudiés, comme celui d'*Arabidopsis thaliana* ou celui du riz (*Oryza sativa*), pour optimiser l'exploitation des informations et des outils existant chez le tournesol afin d'évaluer l'organisation du génome du tournesol et le niveau de synténie avec la plante modèle *Arabidopsis*.

La première étape de ma thèse a donc consisté à définir et à mettre en place la méthodologie qui permet d'exploiter les séquences EST de tournesol qui serviront à analyser l'organisation de son génome. Cette méthodologie utilise des resources informatiques afin d'analyser le grand nombre de séquences et d'informations disponibles. Elle est basée sur la conservation de si-

militudes entre gènes de différentes espèces ainsi que de la structure de ces gènes (découpage intron/exon). Grâce à ces informations, l'exploitation des séquences EST est optimisée et permet de définir des couples d'amorces et des sondes qui servent respectivement à l'amplification génomique afin d'obtenir des marqueurs moléculaires pour l'étude de la macro-structure du génome du tournesol et au criblage d'une banque de clones BAC pour l'étude de la micro-structure. Hongtrakul *et al.* (1998) ont montré que les séquences introniques pouvaient être utilisé comme source de polymorphisme entre variétés de tournesol, nous avons donc choisi de définir des couples d'amorces de part et d'autre des introns en utilisant les informations fournies par l'organisme modèle afin d'optimiser le nombre de marqueurs moléculaires qui pourront être cartographiés. Les sondes définies lors de l'exploitation des séquences EST sont des sondes Overgo. Les sondes Overgo présentent au moins deux avantages comparées aux sondes obtenues par amplification PCR. De par leur taille (40 nucléotides), elles sont beaucoup plus spécifiques et elles peuvent être définies dans une séquence exonique sans avoir le problème du recouvrement avec un intron. Ces sondes ont déjà montré leur efficacité chez les animaux (Cai *et al.*, 1998, Han *et al.*, 2000, Kim *et al.*, 2001, Thomas, *et al.*, 2002) et aussi chez le maïs (Gardiner *et al.*, 2004).

Les résultats fournis par la méthodologie sont ensuite vérifiés expérimentalement en utilisant les séquences EST de tournesol codant la sous-unité S de la rubisco. Ce gène est bien connu chez de nombreuses espèces végétales et permet de vérifier l'efficacité des amorces et des sondes définies à partir des informations fournies par la méthode mise en place.

L'analyse de l'organisation du génome du tournesol et de la synténie avec *Arabidopsis* sont envisagées suivant deux stratégies utilisées avec *a priori* et sans *a priori*.

Avec *a priori*, on part du principe que la microsynténie entre le génome du tournesol et celui d'*Arabidopsis* est conservée. De ce fait, des gènes proches (en distance) chez *Arabidopsis* devraient aussi l'être chez le tournesol. Les séquences EST et ARNm de tournesol similaires à des gènes proches les uns des autres chez *Arabidopsis* devraient aussi être proches chez le tournesol. Donc, des sondes Overgo faites à partir de ces séquences EST et ARNm devraient permettre d'ordonner les clones BAC entre eux, et les couples d'amorces devraient permettre, sous réserve de polymorphisme, de construire une carte génétique qui présentera des marqueurs liés. Les résultats obtenus devraient, de ce fait, confirmer ou infirmer la conservation de microsynténie entre les deux organismes.

Sans *a priori*, on ne tient pas compte de la conservation de microsynténie, seule la conservation des structures entre gènes de différentes espèces nous intéresse. Les sondes Overgo définies à partir de séquences EST ou ARNm de tournesol sont, par définition, présentes sur le génome du tournesol, ce qui nous permettra d'évaluer l'efficacité de la méthode, de même avec les amorces définies à partir de séquences EST et ARNm de tournesol.

En vue de ces objectifs, il a fallu que je combine toutes les informations relatives au génome du tournesol dont je disposais. Il existe trois grands types d'informations sur le génome du tournesol : les banques de clones BAC (dont une disponible au laboratoire (Gentzbittel *et al.*, 2002), les séquences EST (Expressed Sequences Tag) accessibles dans les bases de données publiques (NCBI, EMBL, SRS, etc...) ainsi que celles du laboratoire (banques d'ADNc) et les

cartes génétiques de tournesol de différents croisements parentaux. Nous avons aussi à disposition de nombreux outils bioinformatiques pour l'analyse *in silico* des nombreuses séquences disponibles. Les cartes génétiques vont me permettre, en plus des données de phylogénie, d'avoir des informations sur l'organisation et l'évolution du génome du tournesol ce qui me permettra d'avoir une vision globale de celle-ci (macro-structure). La banque de clones BAC et les séquences EST vont me permettre d'étudier directement les gènes du tournesol et d'avoir une vision plus précise d'une région du génome (micro-structure).

Chapitre 2

Création d'Outils Bioinformatiques

2.1 Iccare : Optimiser l'Exploitation des Séquences EST

2.1.1 Sélectionner les Séquences EST les plus Informatives

L'objectif principal spécifique de l'analyse bioinformatique est de pouvoir définir des couples d'amorces qui serviront à la construction de la cartographie génétique et des sondes Overgo qui serviront à cribler une banque de clones BAC. A partir de séquences EST ou ARNm, les couples d'amorces doivent être définies pour rechercher du polymorphisme, et les sondes Overgo doivent être les plus spécifiques possible pour cribler la banque. Les sondes Overgo doivent être définies dans une région de la séquence qui est très conservée, comme les régions exoniques par exemple. Les couples d'amorces, quant à eux, doivent être très spécifiques des séquences EST ou ARNm pour ne pas amplifier d'autres séquences (notamment les gènes paralogues ou similaires) mais elles doivent aussi permettre l'amplification de régions génomiques polymorphes afin de différentier les cultivars entre eux. L'idéal est donc de définir les couples d'amorces dans des régions exoniques de part et d'autre d'un intron.

L'efficacité d'hybridation et d'amplification des sondes et des couples d'amorces est fonction de deux facteurs : la structure et la qualité de la séquences EST/ARNm qui sert de matrice. Il est indispensable que la qualité de séquençage soit correcte (pour plus d'informations sur les erreurs de séquençage, se reporter à l'annexe *A* page 149), mais cette qualité du séquençage ne peut être vérifiée puisque les profils électrophorétiques de chaque séquence ne sont pas accessibles. Il est aussi indispensable de connaître la structure des séquences EST – de savoir où sont nos régions susceptibles d'être polymorphes, introns, régions UTR 3' et 5', ou bien conservées, exons (pour plus d'informations sur les régions conservées ou polymorphes sur les EST, les ARNm et les gènes se reporter à l'annexe *B* page 151). Les séquences génomiques correspondant à la transcription de nos séquences EST ne sont pas disponibles. Il est donc

impossible de savoir où se trouvent les régions polymorphes ou les régions conservées.

En plus du problème qualitatif, c'est-à-dire de la structure et de la qualité des séquences EST, un problème quantitatif vient s'ajouter : le nombre important de séquences EST et ARNm à analyser. Parmi toutes les séquences EST et ARNm de tournesol disponibles en janvier 2004 (59841 séquences EST et 359 séquences ARNm), il faut pouvoir déterminer quelles sont les séquences qui serviront de matrice à la définition d'amorces et de sondes. La seule façon de pouvoir déterminer quelles sont les séquences les plus intéressantes dans cette foule de données est d'utiliser une méthode informatique pour traiter un maximum de données selon les différents critères mentionnés dans ce paragraphe. Il faut donc tenir compte de la qualité des séquences, avoir accès à leur structure et ce pour l'ensemble des séquences disponibles.

La méthode d'analyse que nous avons choisie consiste donc à comparer les séquences EST ou ARNm à des séquences bien identifiées chez un organisme modèle (*Arabidopsis thaliana*) à l'aide du programme *BLAST*. La seule manière de recueillir des informations sur la structure des séquences EST et ARNm est de les comparer à d'autres séquences bien identifiées comme les ARNm d'un organisme modèle. Toutes les informations sur ces séquences sont disponibles (structure, fonction potentielle, position sur le génome, séquence génomique ou transcript). Il suffit donc de comparer les séquences entre elles puis de transposer les informations de l'organisme modèle aux séquences EST. Ces informations structurales des séquences vont grandement faciliter leur exploitation (identification des régions polymorphes ou conservées).

En plus du transfert d'information, cette méthode permet de vérifier la bonne qualité des séquences (s'il y a trop d'erreurs le score et l'*E*-value du BLAST sont moins bons et la séquence n'est pas sélectionnée). Avec de forte valeur de score et de *E*-value, les séquences EST ou ARNm sont potentiellement orthologues avec le gène de l'organisme modèle et les informations structurales peuvent leur être transposées. Ceci est d'autant plus probable que le score et la *E*-value du BLAST sont élevés.

Cette méthode nous a permis de récupérer 20 407 séquences EST et 214 séquences ARNm qui présentent des similitudes avec des séquences de l'organisme modèle parmi les 59841 séquences EST et les 359 séquences ARNm disponibles au départ. Cette sélection peut ensuite être affinée en définissant des valeurs seuils du score et de l'*E*-value du BLAST plus élevées pour diminuer le nombre de séquences disponibles.

Cette méthode n'est pas parfaite. Certes elle permet d'éliminer les séquences qui présentent des erreurs de séquençage ainsi que les séquences qui sont trop courtes pour être exploitées, mais elle élimine aussi les séquences EST ou ARNm qui sont spécifiques de l'organisme d'intérêt (n'existant pas chez l'organisme modèle) et les séquences qui ont principalement été séquencées dans la partie UTR5' ou UTR3' de l'ARNm (régions très peu conservées) et qui du coup présentent peu ou pas de similitudes avec les séquences de l'organisme modèle. En fonction de ce que l'on désire faire, cette méthode n'est pas toujours la meilleure à utiliser, mais pour ce qui nous conserve, elle est très efficace car elle permet de rapidement trier les séquences EST et ARNm même s'il y a un peu de perte d'information.

La programmation de cet outil informatique a été réalisée en collaboration avec le laboratoire de Génétique Cellulaire de l'INRA de Toulouse, et plus particulièrement avec Thomas Faraut. L'unité de l'INRA a initié le travail en travaillant sur la transposition des informations des gènes de l'homme aux séquences EST d'organismes animaux d'intérêt (poulet, porc, cheval, etc). Tout en respectant les principes d'analyses énoncés ci-dessus, j'ai adapté une partie des programmes déjà élaborés pour qu'ils fonctionnent avec des plantes (organismes modèles : *Arabidopsis thaliana*, organismes d'intérêts : tournesol, riz, maïs, etc). Une fois mise au point l'ensemble des programmes a été intégré au sein d'un même software élaboré de façon à traiter les séquences EST ou ARNm de n'importe quel organisme (aussi bien animaux que végétaux) et de les comparer automatiquement aux séquences codantes d'un organisme modèle (*Arabidopsis thaliana* pour les plantes et *Homo sapiens* pour les animaux). Cet outil permet donc d'analyser rapidement et automatiquement un grand nombre de séquences de différents organismes dit "d'intérêt" contre les séquences codantes d'un organisme modèle.

Cet outil s'appelle Iccare pour "Interspecific Comparative Clustering and Annotation foR Est" (Regroupement par comparaison interspécifique et annotation des EST). Le server web Iccare est un outil simple et efficace pour l'annotation spécifique des EST pour les approches de génomique comparée. Iccare utilise toutes les séquences EST et ARNm d'un organisme d'intérêt dans les bases de données publiques et les compare aux séquences transcrites d'un organisme de référence (en l'occurence *Homo sapiens* ou *Arabidopsis thaliana*). Les résultats sont graphiquement présentés en fonction des localisations des gènes sur les chromosomes de l'organisme modèle. L'information de structure des gènes et l'information de similitudes entre séquences sont combinées et représentées graphiquement afin de faciliter la détermination de la nature des séquences de l'organisme d'intérêt. L'utilisateur peut alors désigner des couples d'amorces ou des sondes Overgo pour faire de la cartographie génétique ou physique.

2.1.2 Iccare, un Server Web Efficace

Fonctionnement de Iccare

Iccare est composé de deux grandes parties : une partie "non-visible" par l'utilisateur qui constitue toute la partie analytique (traitement des données brutes) et une partie "visible" par l'utilisateur qui présente l'ensemble des résultats (interface web).

La première partie sert à la récupération des données, au formatage de celles-ci, à la recherche de similitudes entre séquences puis au formatage des fichiers de résultats ; la deuxième partie présente les résultats contenus dans les fichiers formatés dans un site web interactif qui permet la représentation graphique des résultats tout en combinant différents outils pour l'exploitation des séquences par l'utilisateur.

La première partie – partie analytique – est programmée en script csh et en script PERL (avec les modules PERL : :Seqlo et PERL : :DB). Elle est elle-même divisée en trois étapes : la récupération des informations, la recherche de similitudes et la préparation des fichiers de

résultats (Figure 2.1).

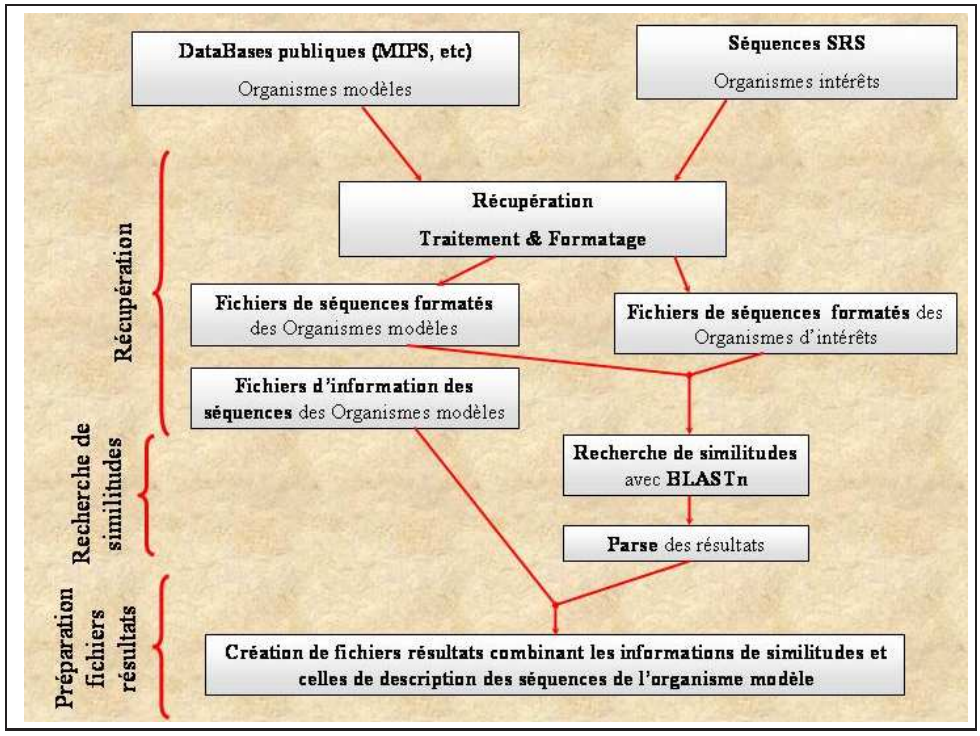


FIG. 2.1 – Architecture de la partie analytique de Iccare.

Dans un premier temps, la récupération des séquences et des informations relatives à celles-ci est effectuée pour les organismes modèles (*Homo sapiens*, *Arabidopsis thaliana*) puis pour les organismes d'intérêt. Pour les organimes modèles, plusieurs types de fichiers sont récupérés : fichiers de séquences, fichiers contenant les informations relatives aux différents gènes et les fichiers relatifs aux différents chromosomes de l'organisme modèle. Les fichiers de séquences contiennent soit les séquences entières des ARNm (pour l'homme) soit uniquement la région codante de l'ARNm (c'est-à-dire la région traduite, pour *Arabidopsis*) des différents gènes au format FASTA. Pour l'homme, ce fichier provient de l'UniGene du National Center for Biotechnology Information (`http://www.ncbi.nlm.nih.gov`). Pour *Arabidopsis*, ce fichier provient du Munich Information center for Protein Sequences (MIPS, `http://mips.gsf.de`; `ftp://ftpmips.gsf.de/cress/arabidna/arabi_cds_v090704.gz`). Les fichiers contenant les informations relatives aux gènes permettent de savoir sur quel chromosome sont localisés ces gènes, leur fonction confirmée ou potentielle, leur structure (épissage intron/exon) et leur orientation sur l'ADN (sens/anti-sens). Pour *Arabidopsis*, ces informations sont contenues dans 7 fichiers (un par chromosome plus le chloroplaste et la mitochondrie) sur le site du MIPS ; ces fichiers sont du type "chromo1_anno_v090704.dat.gz". Pour l'Homme, ces fichiers sont récupérés à partir du site Ensembl (`http://www.ensembl.org`). Le troisième type de fichier contient la taille des différents chromosomes des organismes modèles (plus les positions des bandes cytologiques pour l'homme). L'ensemble de ces fichiers est ensuite formaté de manière à faciliter le traitement de l'information. Pour les séquences EST et ARNm

des différents organismes d'intérêt (et en particulier *Helianthus annuus*), les séquences sont récupérées au format FASTA dans la base de données SRS du Génopôle de l'INRA de Toulouse. Ces séquences subissent un masquage des régions contenant des vecteurs animaux et végétaux avec UniVec (ftp://ftp.ncbi.nih.gov/pub/UniVec) ainsi qu'un masquage des régions répétées connues chez les animaux et les plantes avec RepeatMasker (A. Smit and P. Green, http://ftp.genome.washington.edu/RM/RepeatMasker.html). Puis, les séquences sont formatées et indexées pour faciliter la récupération d'une séquence en particulier.

La deuxième étape consiste ensuite, après avoir récupéré toutes les séquences, à lancer la recherche de similitudes à l'aide de l'option blastn du programme BLAST (Altschul *et al.*, 1997). Les séquences EST et ARNm d'un organisme d'intérêt sont comparées à toutes les séquences codantes d'un organisme modèle (*Arabidopsis thaliana* pour les plantes et *Homo sapiens* pour les animaux). Les résultats du BLAST sont filtrés en fonction de la valeur de l'expect value, notée *E*-value. Cette valeur est normalisée afin d'avoir des *E*-value comparables d'un blast à l'autre (normalisation équivalente à une base de données d'un million de résidus). Seules les similarités – ou plus précisément, en utilisant la terminologie de BLAST, Highest Scoring Pairs (HSP, score d'identité le plus élevé) – avec une *E*-value $< 10^{-5}$ sont récupérées (plus l'*E*-value est proche de zéro moins la probabilité de similitude est due au hasard). Afin d'éviter le stockage des fichiers blast, qui sont très gros même compressés (pouvant aller de 3 Mo à 100 Mo en fonction des organismes comparés), le fichier de résultat blast est traité de manière à fortement résumer l'information. Ensuite, les résultats de ce fichier ainsi que les identifiants de toutes les séquences utilisées par le blast (séquences organisme modèle et séquences organisme d'intérêt) sont inserés dans une base de données de type MySQL. Cette base de données va permettre de créer un fichier dans lequel le gène de l'organisme modèle qui présente la plus forte similitude (soit la valeur d'*E*-value la plus faible) sera affecté à chaque séquence EST ou ARNm de l'organisme d'intérêt (aucun gène ne sera affecté s'il n'y a pas de similitude). Ainsi à chaque EST ou ARNm est associé un seul et unique gène de l'organisme modèle, mais à chaque gène de l'organisme modèle peuvent être associés plusieurs séquences EST ou ARNm de l'organisme d'intérêt. Ceci permet de regrouper les séquences EST qui pourraient être issues d'un seul et même ARNm (les EST étant des fragments partiels d'ARNm).

La dernière étape consiste à combiner les informations relatives aux séquences de l'organisme modèle aux résultats de recherche de similitudes en créant un fichier unique contenant l'ensemble de l'information. Celui-ci servira à la deuxième partie de Iccare – site web – pour la représentation de ces résultats. Une fois cette étape terminée, tout est opérationnel pour le fonctionnement du site web.

La deuxième partie de Iccare est programmée en PERL (avec module PERL : :CGI et le module PERL : :GD) ainsi qu'en PHP et en HTML et associé à des scripts en JavaScript et en CSS. Les scripts JavaScript sont inserés dans tous les programmes du site web (scripts PERL : :CGI, PHP ou HTML), ils permettent de rendre la page web dynamique et de créer une interaction avec l'utilisateur. Iccare est hébergé par le Génopôle Toulouse Midi-Pyrénées (France) et est accessible à l'adresse suivante : http://genopole.toulouse.inra.fr/bioinfo/Iccare).

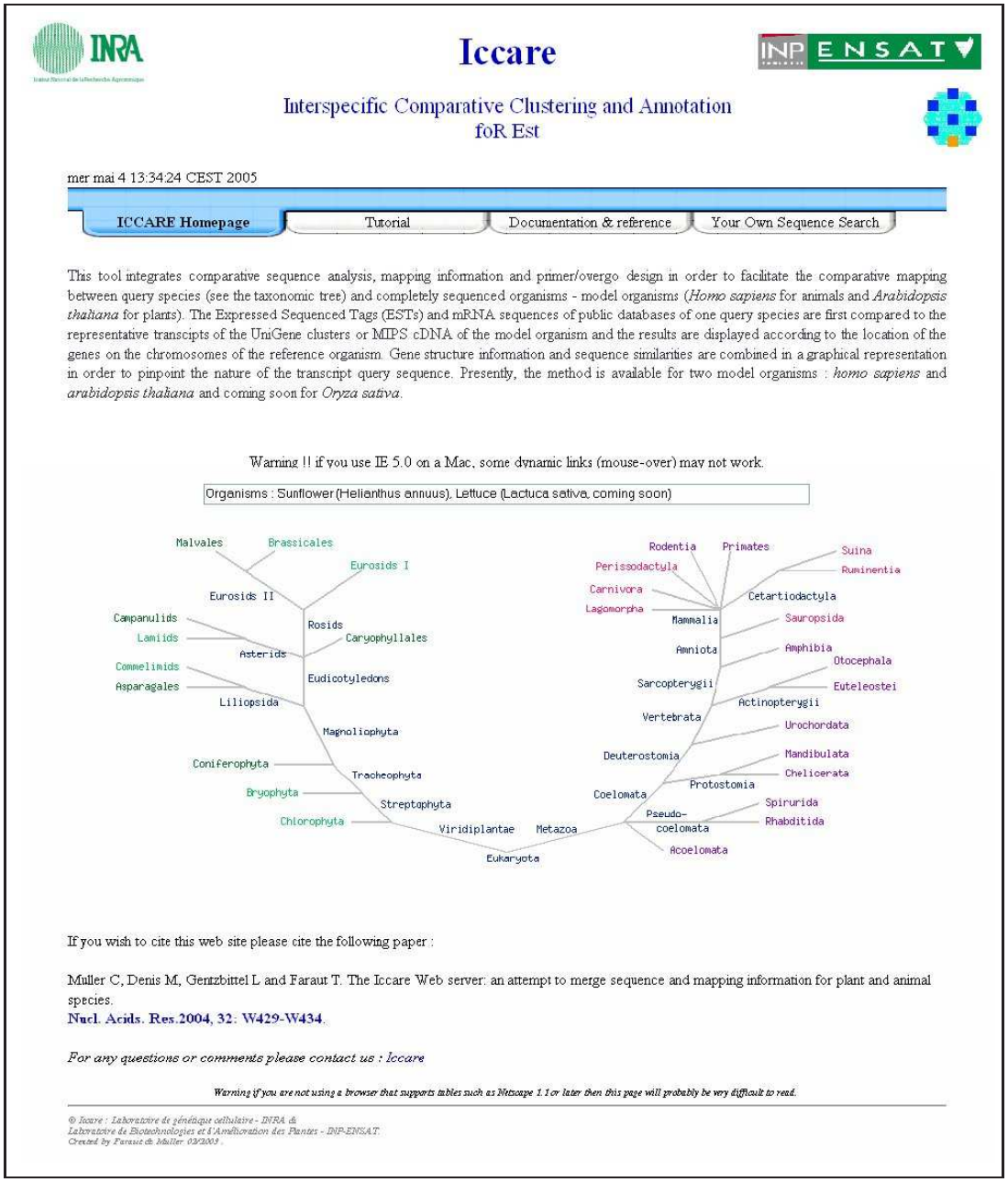


FIG. 2.2 – Page d'accueil du site web Iccare.

Description du site web Iccare

La page d'accueil du site contient deux types d'informations (Figure 2.2). Le premier type regroupe les informations relatives aux processus de fonctionnement de Iccare, toute la documentation, le didacticiel ainsi que les références qu'utilise Iccare (signet *Documentation* et *Tutorial*). Le second correspond aux informations relatives aux résultats pré-calculés de Iccare ainsi qu'aux résultats issus de la soumission de séquence(s) d'un utilisateur.

Les résultats pré-calculés d'Iccare sont accessible par l'intermédiaire de l'arbre taxonomique contenant l'ensemble des organismes d'intérêt disponible sur la page d'accueil ou par le signet *Your Own Sequence Search* pour des séquences personnelles. La partie intitulée *Your Own Sequence Search* permet de soumettre ses propres séquences à Iccare qui traitera ces séquences de la même façon que les séquences des organismes déjà disponibles grâce à une automatisation de la partie analytique. Les résultats de cette soumission seront ensuite présentés de façon identique aux résultats pré-calculés de Iccare. Pour faciliter la navigation sur le site, se référer à la partie Didacticiel du site (*Tutorial*).

Iccare permet la comparaison entre un organisme d'intérêt et un organisme modèle, soit animal soit végétal, et il existe très peu de différences entre la partie animale et la partie végétale car la totalité des programmes a été créé afin de fonctionner avec les deux, mais pour plus d'informations sur la partie animale reportez-vous à la publication (Muller *et al.*, 2004) dans l'annexe *C*.

Pour illustrer le site web Iccare, je n'utiliserai que la partie végétale en prenant l'exemple du tournesol (mais cette présentation serait la même avec un autre organisme végétal ou animal). La *page d'accueil* permet de sélectionner un organisme végétal qui a été comparé à *Arabidopsis* ou un organisme animal qui a été comparé à *Homo sapiens*. Une fois l'organisme sélectionné, en l'occurrence le tournesol (phylum : lamiids), l'ensemble des chromosomes (caryotype) de l'organisme modèle est présenté (Figure 2.3). Sur cette page, l'utilisateur peut sélectionner un chromosome, en cliquant directement dessus, ou bien rechercher un gène de l'organisme modèle ou une EST de l'organisme d'intérêt.

La page suivante présente le chromosome sélectionné de l'organisme modèle ainsi que la répartition des gènes sur ce chromosome (Figure 2.4). A certains grossissements, tous les gènes ne peuvent être représentés alors la priorité est donnée aux gènes qui présentent des similitudes avec des EST de l'organisme d'intérêt. Les boutons à gauche de l'image permettent aux utilisateurs de se déplacer et de zoomer le long du chromosome. Lorsque l'agrandissement est suffisant, l'utilisateur peut distinguer tous les gènes repartis sur le chromosome. Les gènes qui présentent des similitudes avec des EST de l'organisme d'intérêt sont représentés par un identifiant et un rectangle vert, en revanche ceux ne présentant pas de similitudes sont représentés par un identifiant en noir et un rectangle bleu. En passant la souris sur les identifiants des gènes, une boîte de dialogue apparaît contenant un lien vers la page de *visualisation des résultats*.

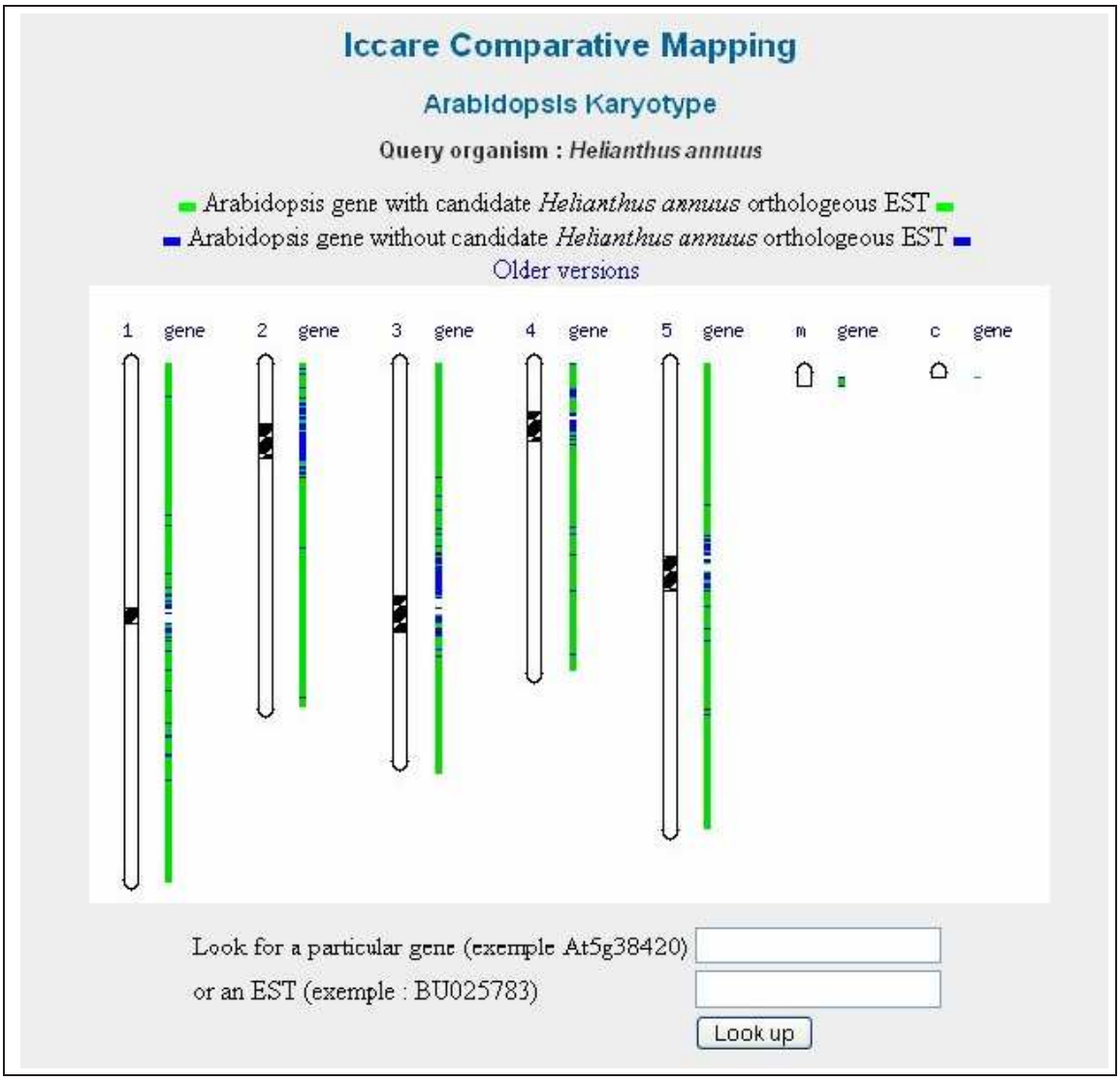


FIG. 2.3 – Représentation du caryotype d'*A. thaliana*.

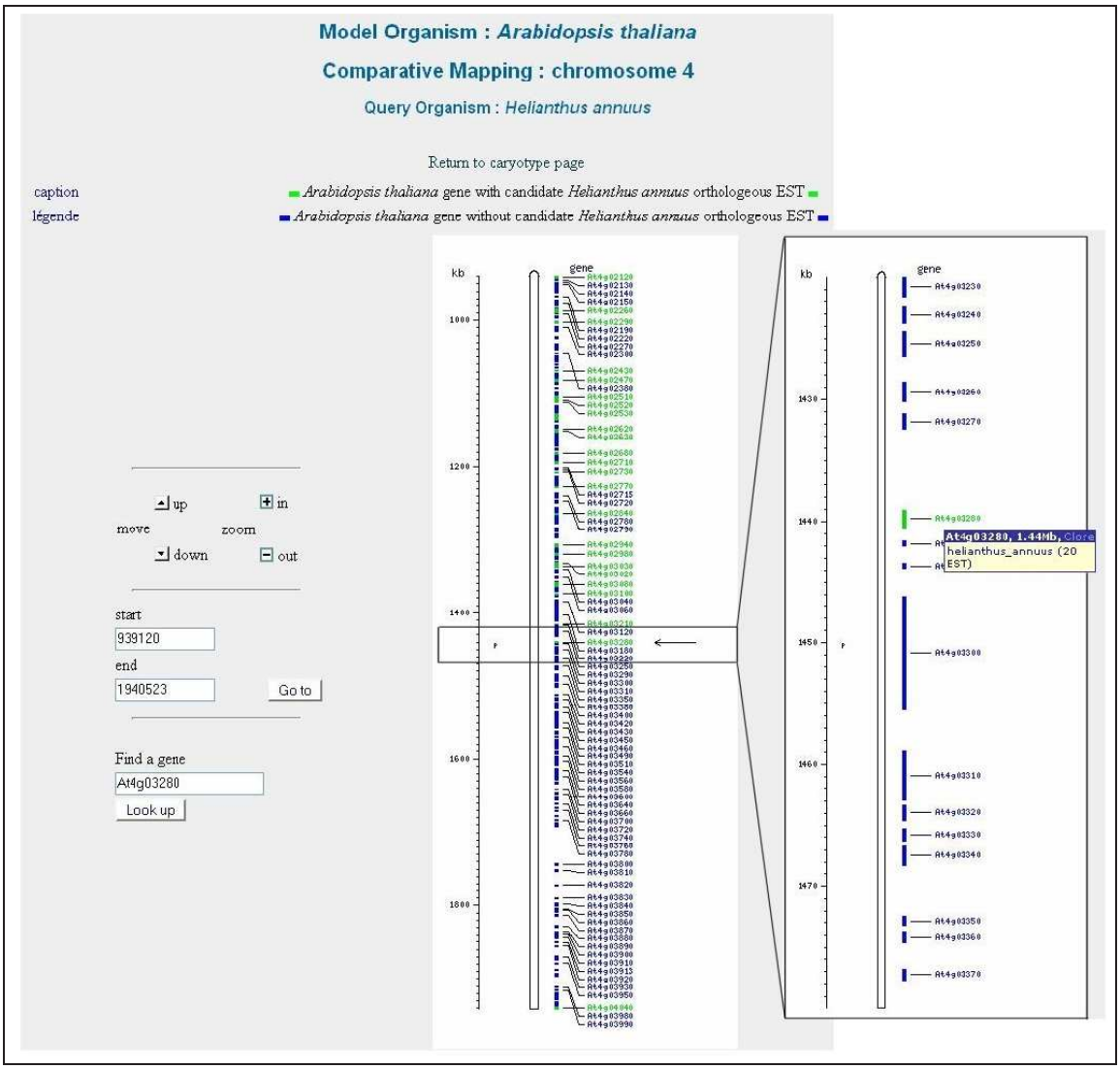


FIG. 2.4 – Représentation et répartition des gènes du chromosome 4 d'*A. thaliana*.

Model organism gene At4g03280

Iccare version: 22 janvier 2004 (UniGene: 20_mars_2003)

Query organism: *Helianthus annuus*

légende caption

Back to chromosomal map

Cluster	Gene	definition	Contig	GoldenPath	Ensembl	MapView
At4g03280	?	AT4g03280/F4C21_21	F4C21	?	?	At4g03280

Repesentative sequence of the cluster					Mapping information (goldenpath)			
Sequence	**Genbank ID**	**length**	**CDS**	**LocusId**	**chromosome**	**physical (Mb)**	**interval (Mb)**	**predicted cytogenetic**
At4g03280	?	690	(1,691)		4	1.440	1.439-1.441	

Graphical representation of sequence similarities :alignment

Graphical representation of genome evolution of Arabidopsis :duplication

Multi-Alignment representation of sequences :multi-species alignment

Homologous EST

EST	length	score (e-value)	HSP : Est[start-end] --> Mod[start-end] (length, orientation, e-value,% identity)	2e hit	blastn	tblastx
CD847222	616	206(5e-52)	[99-517]-->[268-689](422, Plus/Plus, 206, 1e-52, 81%):	(>e-5)	blast	tblastx
CD858421	422	192(6e-48)	[43-422]-->[307-689](383, Plus/Minus, 192, 2e-48, 81%):	(>e-5)	blast	tblastx
CD857955	422	174(1e-42)	[43-413]-->[316-689](374, Plus/Minus, 174, 4e-43, 81%):	(>e-5)	blast	tblastx
CD857407	422	174(1e-42)	[43-413]-->[316-689](374, Plus/Minus, 174, 4e-43, 81%):	(>e-5)	blast	tblastx
CD857796	422	174(1e-42)	[43-413]-->[316-689](374, Plus/Minus, 174, 4e-43, 81%):	(>e-5)	blast	tblastx
CD858138	422	174(1e-42)	[43-413]-->[316-689](374, Plus/Minus, 174, 4e-43, 81%):	(>e-5)	blast	tblastx
CD857878	422	174(1e-42)	[43-413]-->[316-689](374, Plus/Minus, 174, 4e-43, 81%):	(>e-5)	blast	tblastx
CD857267	422	174(1e-42)	[43-413]-->[316-689](374, Plus/Minus, 174, 4e-43, 81%):	(>e-5)	blast	tblastx
CD858323	422	174(1e-42)	[43-413]-->[316-689](374, Plus/Minus, 174, 4e-43, 81%):	(>e-5)	blast	tblastx
CD858350	396	172(6e-42)	[43-380]-->[349-689](341, Plus/Minus, 172, 1e-42, 81%):	(>e-5)	blast	tblastx
CD846369	705	170(4e-41)	[262-611]-->[268-617](350, Plus/Plus, 170, 6e-42, 81%):	(>e-5)	blast	tblastx
CD857208	422	168(1e-40)	[43-422]-->[307-689](383, Plus/Minus, 168, 2e-41, 80%):	(>e-5)	blast	tblastx
CD857261	422	167(2e-40)	[43-413]-->[316-689](374, Plus/Minus, 167, 9e-41, 80%):	(>e-5)	blast	tblastx
CD857277	422	167(2e-40)	[43-413]-->[316-689](374, Plus/Minus, 167, 9e-41, 80%):	(>e-5)	blast	tblastx
CD856895	360	163(3e-39)	[43-331]-->[398-689](292, Plus/Minus, 163, 1e-39, 82%):	(>e-5)	blast	tblastx
CD858411	284	117(1e-25)	[43-284]-->[445-689](245, Plus/Minus, 117, 7e-26, 81%):	(>e-5)	blast	tblastx
CD857453	272	117(1e-25)	[43-272]-->[457-689](233, Plus/Minus, 117, 7e-26, 81%):	(>e-5)	blast	tblastx
CD858407	284	117(1e-25)	[43-284]-->[445-689](245, Plus/Minus, 117, 7e-26, 81%):	(>e-5)	blast	tblastx
BU035917	246	71(1e-11)	[23-78]-->[634-689](56, Plus/Plus, 71.9, 4e-12, 91%):	(>e-5)	blast	tblastx
BQ973368	287	58(9e-08)	[40-96]-->[634-690](57, Plus/Plus, 58.0, 6e-08, 87%):	(>e-5)	blast	tblastx

(Only similarities with an e-value beneath e-5 are considered)

FIG. 2.5 – Page de description des résultats de similitudes pour le gène At4g03280 d'*A. thaliana*.

Avant d'avoir accès à la visualisation des résultats sous forme graphique, ceux-ci sont d'abord présentés sous forme de tableaux de synthèse (Figure 2.5). Dans la partie supérieure de la page, l'utilisateur retrouve une rapide description du gène (fonction potentielle, taille de la région codante, numéro de chromosome et emplacement sur ce chromosome). Plusieurs liens permettent d'obtenir de plus amples informations sur ce gène en redirigeant l'utilisateur sur le site du MIPS (d'où proviennent les séquences codantes d'*Arabidopsis*) ainsi que du NCBI pour ce qui est du clone BAC et de MapView. Dans la partie basse de la page, l'utilisateur retrouve un tableau contenant l'ensemble des séquences EST ou ARNm de l'organisme d'intérêt qui présentent des similitudes avec ce gène. Ce tableau contient la taille des séquences, un résumé du blast (*E*-value, score et HSP) et un lien vers l'option blastn et/ou tblastx du BLAST. Ce blastn (nucléotides contre nucléotides) et ce tblastx (traduction des 6 cadres de lecture d'une séquence nucléotidique contre la traduction des 6 cadres de lecture d'une autre séquence nucléotidique) permettent de voir si la séquence EST a des similitudes avec d'autres gènes de l'organisme modèle que celui indiqué sur la page de résultats (Figure 2.6).

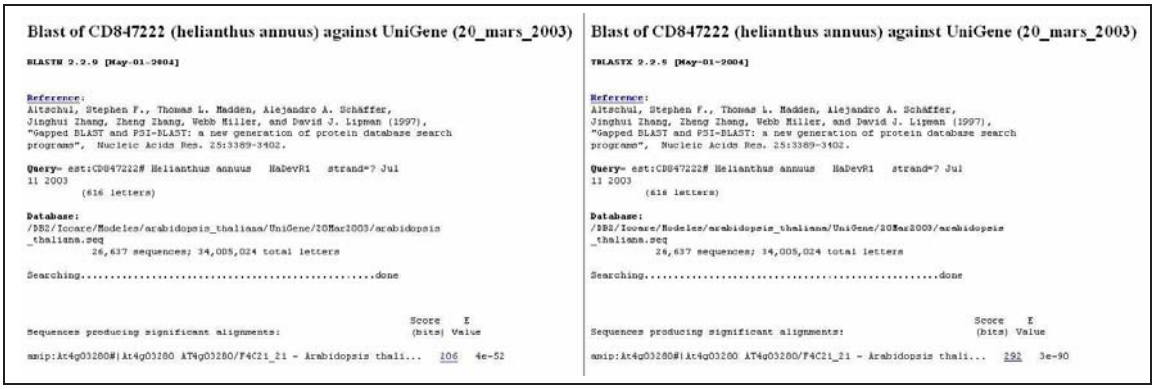


FIG. 2.6 – Résultats du blastn (à gauche) et du tbalstx (à droite) de la séquence CD847222 du tournesol contre l'ensemble des séquences codantes d'*A. thaliana*.

Au centre de la page se trouvent trois liens : *Alignment*, *Multi-species alignment* et *duplication*. Les liens *multi-species alignment* et *duplication* n'étaient pas inclus dans la version publiée de Iccare, ils ont été ajoutés par la suite. L'alignement multiple sera décrit dans le paragraphe intitulé "Amélioration de Iccare", page 47, et le lien *duplication* permet d'accéder au site appelé "Synteny Search" décrit plus loin dans ce manuscrit (paragraphe "Exploiter la Synténie d'autres Organismes", page 81). Le lien *alignment* permet de visualiser sous forme graphique les similitudes entre les séquences EST de l'organisme d'intérêt et le gène de l'organisme modèle (Figure 2.7 partie A).

Dans le cadre le plus haut se trouvent toutes les informations relatives à la structure du gène At4g03280. La première ligne correspond à l'ADN génomique, les rectangles jaunes (dans l'orientation anti-sens [sens −] de l'ADN ou verts s'ils sont dans l'orientation sens [sens +]) correspondent aux exons et les chiffres entre les rectangles correspondent à la taille des introns. La deuxième ligne correspond à la concaténation des exons, mimant ainsi le processus de transcription qui permet l'obtention des ARNm. Les différents exons sont représentés par

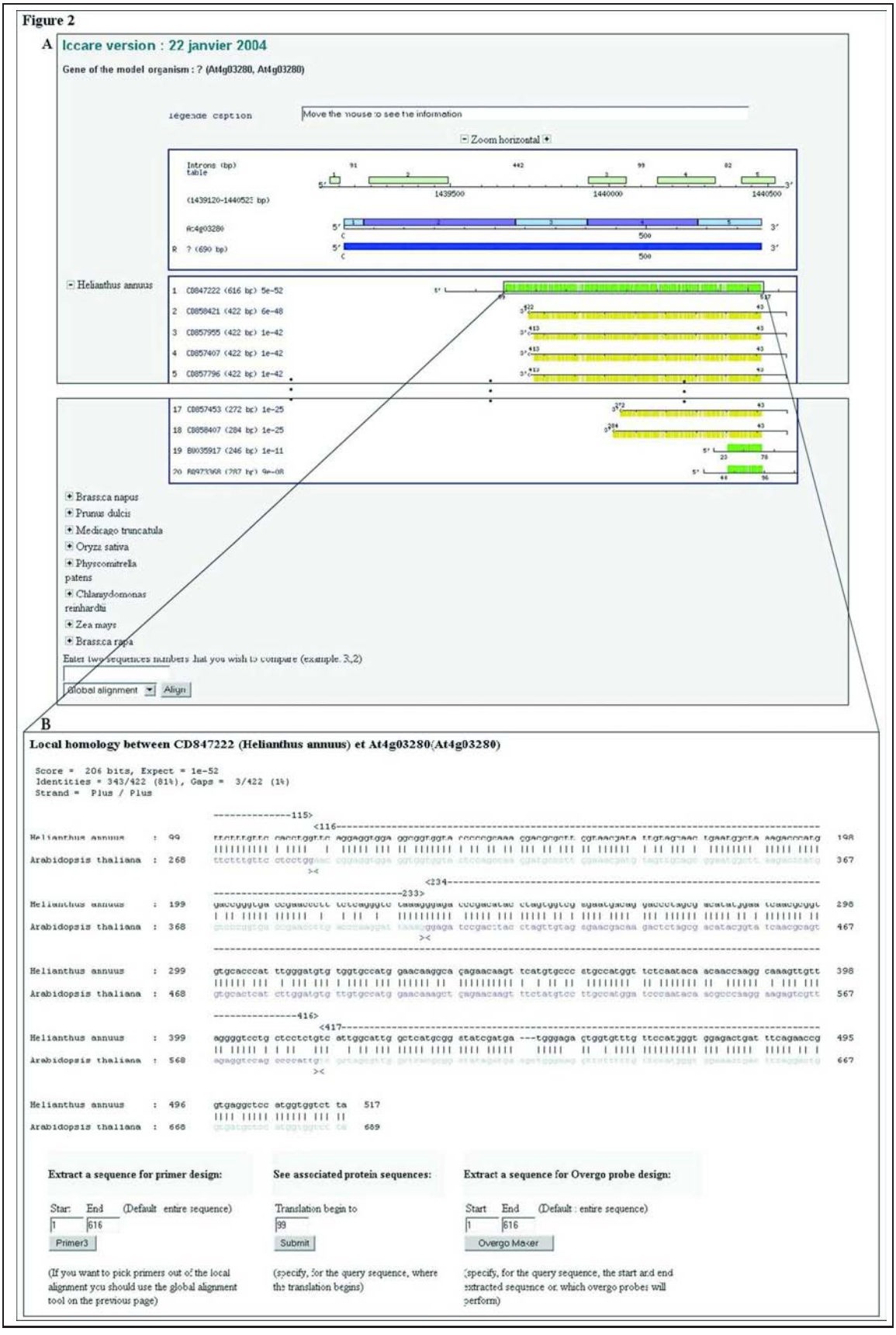


FIG. 2.7 – Représentation graphique de la structure du gène At4g03280 et de l'alignement local.

une alternance de couleurs (bleu–violet). Et enfin, la troisième ligne correspond à la région traduite de l'ARNm (la partie qui sert de matrice à la synthèse de la protéine). Le cadre, en dessous, correspond à la représentation des séquences EST ou ARNm de l'organisme d'intérêt. Dans la partie gauche de ce cadre se trouve les identifiants des séquences EST avec leur taille et l'*E*-value du blast et dans sa partie droite se trouve la représentation graphique de la séquence EST et des similitudes. La séquence est symbolisée par la ligne noire et les similitudes sont symbolisées par un rectangle vert ou jaune (reverse–complémentée). La séquence EST et les similitudes sont alignées avec la séquence du gène de l'organisme modèle ce qui permet de transposer les informations de l'organisme modèle aux séquences EST (positions des introns, des UTR 5' et 3'). En dessous des deux cadres se trouve une liste d'autres organismes qui permet en cliquant sur le signe +, à côté de chaque nom d'espèce, de faire apparaître les séquences EST de cet organisme et de comparer tous les résultats. Chaque rectangle vert ou jaune sur les séquences EST est un lien dynamique vers l'alignement des séquences (Figure 2.7 partie B). La séquence de l'EST sélectionnée (CD847222) est alignée avec la séquence codante du gène de l'organisme modèle (At4g03280) pour la région présentant des similitudes entre ces deux séquences. En plus de l'alignement local de ces séquences, les informations sur la struture du gène At4g03280 y sont combinées. Le haut de la page contient toutes les informations relatives au *BLAST* (scrore, *E*-value, sens de l'alignement, etc.). Au centre se trouve l'alignement local des séquences avec la séquence de l'EST en noir alignée avec la séquence codante du gène en bleu et violet (correspondant aux différents exons). Enfin, en bas de la page se trouvent trois outils qui permettent de travailler sur les séquences. Celui du centre permet de refaire l'alignement local des séquences mais cette fois-ci au niveau protéique et non plus au niveau nucléotidique. Les deux autres outils permettent de désigner des amorces ou des sondes Overgo à partir de la séquence EST à l'aide respectivement de Primer3 (Rozen et Skaletsky, 2000) et de Overgo Maker 40 (http://www.genome.wustl.edu).

2.1.3 Améliorations apportées à Iccare depuis la Publication

De nouveaux outils ont été ajoutés comme l'alignement multiple de séquences, la recherche de gènes dupliqués ou appartenant à des familles multigéniques. Un suivi de la fréquentation du site est aussi possible. Et au niveau analytique, des procédures d'automatisation des mises à jour et de l'ajout de nouveaux organismes sont en cours.

Ajouts de nouveaux outils

Depuis la publication de l'article, en juillet 2004 (voir annexe *C* page 154), de nouveaux outils ont été ajoutés à Iccare : l'alignement multiple et la recherche de duplication ou de famille multigénique chez les organismes végétaux modèles. La recherche de gènes dupliqués ou appartenant à une famille multigénique est une information capitale dans le cas de notre étude. Savoir que le gène utilisé comme matrice est unique ou pas donne une indication sur

les résultats que l'on aura après amplification ou après hybridation. Si le gène matrice est dupliqué ou appartient à une famille multigénique les résultats d'amplification ou d'hybridation peuvent laisser présager un nombre de bandes ou un nombre de clones positifs plus important que si le gène matrice est unique. Cette information est importante et permet de faciliter l'analyse des résultats expérimentaux. Cette partie sur la recherche de duplication est développée plus loin dans le paragraphe intitulé "Exploiter la Synténie d'autres Organismes" page 81. Le deuxième outil mis en place est le couplage de Iccare au site d'alignement multiple Multalin (Corpet, 1988, http://prodes.toulouse.inra.fr/multalin/multalin.html). Cette association permet de faire un alignement multiple du gène de l'organisme modèle avec toutes les séquences (ou seulement les séquences sélectionnées) d'un ou de plusieurs des organismes d'intérêt présents sur Iccare afin de rechercher les régions conservées ou les polymorphismes spécifiques de certains organismes ou familles d'organismes. Cette version de Multalin associée à Iccare est totalement identique à la version publiée pour ce qui est du traitement des données (algorythme d'alignement). Par contre, la présentation des résultats de la version couplée à Iccare a été légèrement modifiée par rapport à la version originale.

La version originale de Multalin permet de mettre en rouge les nucléotides communs à au moins 90% des séquences alignées et en bleu ceux communs à au moins 50%. Cette présentation permet de faciliter la visualisation des régions conservées entre séquences. Ce pourcentage est calculé en divisant le nombre de nucléotides communs majoritaires par le nombre de séquences totales. Cette méthode de calcul est efficace lorsque l'on compare un ensemble de séquences de même taille. Mais lorsque les séquences comparées n'ont pas la même taille, ce qui est le cas lorsque l'on compare des séquences codantes de gènes à des EST (qui sont des segments partiels du côté 5' ou 3' de l'ARNm), cette méthode de calcul ne permet pas de visualiser correctement les régions conservées entre séquences. Le calcul ne tient pas compte du fait que les séquences en 3' n'ont pas la partie 5', et réciproquement. De ce fait, il ne faut pas tenir compte de cette absence de séquences car ce n'est pas une absence de similitudes. Puisque Iccare est basé sur la comparaison des séquences codantes de gènes avec des séquences EST, il est important de tenir compte de ce mode de présentation. Il a donc fallu modifier la méthode de calcul du pourcentage pour améliorer la présentation des résultats de Multalin. La version modifiée de Multalin calcule le pourcentage uniquement en fonction des séquences présentes dans la région comparée (pour plus d'informations sur les modifications de la représentation des résultats de Multalin pour Iccare, reportez-vous à l'annexe *D* page 161). Cette nouvelle présentation des résultats permet de plus facilement repérer les régions conservées entre séquences même avec des séquences EST séquencées en 5' et/ou en 3'.

Mise à jour

Depuis le mois d'Octobre, le site est régulièrement utilisé. La grande majorité des pages consultées concerne les animaux et seuls 10% concernent les plantes. La partie plante n'est pas aussi utilisée que la partie animale (bien sûr, je ne comptabilise pas ma propre utilisation de Iccare). Cette différence est peut être due au fait que la recherche de synténie est déjà bien plus utilisée chez les animaux qu'elle ne l'est chez les végétaux. En plus, la partie animale

possède des informations supplémentaires par rapport à la partie végétale. La partie animale utilise les séquences codantes des gènes d'*Homo sapiens* comme référence mais en plus pour chacun de ces gènes, l'information d'orthologie avec les gènes de souris (*Mus musculus*) est aussi présentée alors que pour les plantes ce type d'information n'est pas encore accessible.

La partie soumission de séquences personnelles est régulièrement utilisée. Au fil des mois, certains instituts sont devenus des utilisateurs réguliers. Bien entendu, le plus gros utilisateur de la partie animale reste l'INRA de Toulouse (environ 30%), mais d'autres INRA (Jouy et Rennes) l'utilisent aussi ainsi que le NIAS (National Institut of Agrobiological Sciences) au Japon, par exemple. Le NIAS utilise aussi de temps en temps la partie végétale, mais à part moi, l'utilisation de cette partie est moins fréquente. On y trouve cependant quelques utilisateurs d'Euralis et d'Infobiogen. Sa fréquentation m'incite à penser que le site Iccare est utile et qu'il est utilisé, il faut donc le mettre à jour régulièrement en essayant d'y apporter de constantes nouveautés (nouveaux outils intégrés, nouveaux organismes, etc.).

Iccare n'a pas été mis à jour depuis la publication. Mais avant de mettre à jour tous les organismes déjà présents, nous souhaitons d'abord réussir à parfaitement automatiser les différentes étapes de cette mise à jour (la récupération des séquences, leur traitement, les comparaisons et la mise en place sur le site). En plus de ces futures mises à jour, nous souhaitons ajouter de nouveaux organismes d'intérêt pour compléter les différentes branches taxonomiques. Pour rendre la recherche de comparaisons encore plus performante, nous souhaitons aussi ajouter deux nouveaux organismes modèles, *Danio Rerio* pour les poissons et *Oryza sativa* pour les monocotylédones. Ceci permettra de créer deux nouvelles branches dans l'arbre des comparaisons qui seront plus spécifiques des différents organismes d'intérêt. Une fois ces modifications effectuées, Iccare se divisera en quatre branches. La première branche représentera les plantes monocotylédones et contiendra l'organisme modèle *Oryza sativa* et les organismes d'intérêt comme le maïs, le blé, le sorgho, la canne à sucre, etc. La deuxième branche représentera les plantes dicotylédones et contiendra l'organisme modèle *Arabidopsis thaliana* et des organismes d'intérêt comme le tournesol, la laitue, le soja, la luzerne, le lotier, le prunier, la tomate, la pomme de terre, etc. La troisième branche représentera les animaux supérieurs et contiendra l'organisme modèle *Homo sapiens* et des organismes d'intérêt comme la souris, le rat, le boeuf, le porc, le cheval, le lapin et d'autres. Enfin, la dernière branche représentera les poissons et contiendra l'organisme modèle *Danio rerio* et les organismes d'intérêt suivants le saumon, la truite.

2.1.4 Résultats Généraux chez les Plantes

Relations taxonomiques entre les organismes d'intérêt et la plante modèle

Lors de la mise en ligne de Iccare, neuf organismes végétaux d'intérêt étaient disponibles. Une fois toutes les séquences EST et ARNm traitées et analysées par Iccare, je me suis intéressé

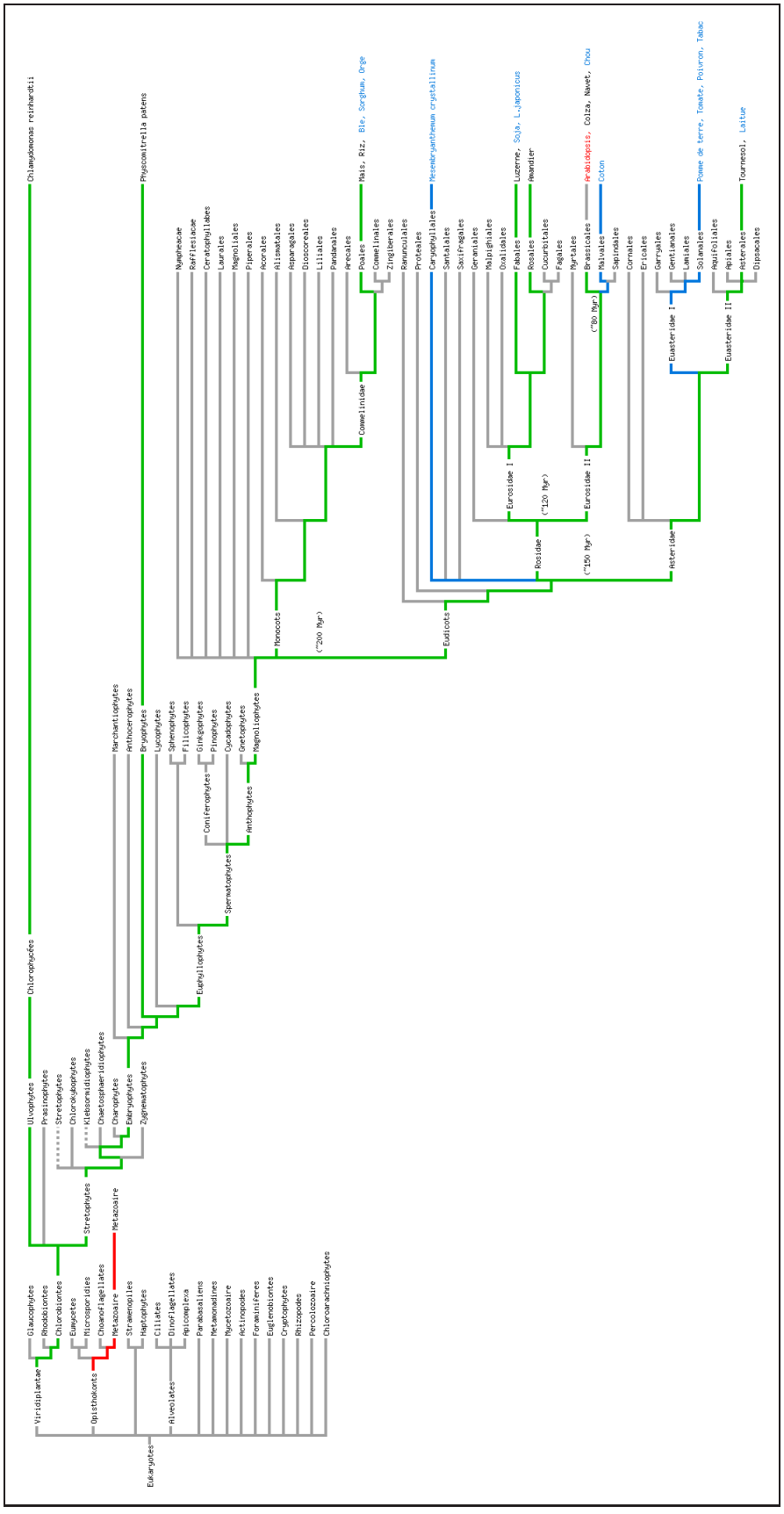


FIG. 2.8 – Arbre taxonomique des organismes eucaryotes végétaux, synthèse des données fournies par le NCBI, The Tree of Life, Museum of Paleontology, Lecointre et Le Guyader, 2001 et Guignard, 2001.

aux résultats obtenus pour la recherche de similitudes en fonction de l'organisme utilisé et de son éloignement phylogénétique avec la plante modèle, *Arabidopsis thaliana*. Les 9 espèces végétales utilisées sont plus ou moins proches phylogénétiquement de l'espèce modèle (Figure 2.8). Cet arbre taxonomique non exhaustif a été construit en combinant les informations de 2 livres (" Classification phylogénétique du vivant ", Lecointre et Le Guyader, 2001 et " Botanique - Systématique moléculaire ", Guignard, 2001) et de 3 sites web : " The Tree of Life " (http://www.tolweb.org/), " Museum of Paleontology " de l'université de Californie, Berkeley (http://www.ucmp.berkeley.edu/) et le " TaxBrowser " du National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/).

Le végétal le plus éloigné de la plante modèle est une algue verte unicellulaire, *Chlamydomonas reinhardtii* [Cre], qui fait partie des Chlorophycées. Ensuite, la mousse, *Physcomitrella patens* [Ppa], est utilisée pour représenter les plantes Bryophytes (plantes qui n'ont pas de fleur). Chez les plantes à fleurs (angiospermes) qui se divisent en deux branches, les monocotylédones (Liliopsida) et les dicotylédones (Eudicots), plusieurs représentantes sont utilisées. Le riz (*Oryza sativa* [Osa]) et le maïs (*Zea mays* [Zma]) représentent les monocotylédones. Du côté des dicotylédones, les plantes utilisées appartiennent à deux divisions différentes : les Asterids, avec le tournesol (*Helianthus annuus* [Han]), et les Rosids. Le groupe des Rosids se subdivisent en deux branches, les Eurosids I et les Eurosids II. Au sein des Eurosids II, deux espèces végétales représentent chacune un genre différent ; l'amandier (*Prunus dulcis* [Pdu]) du genre Rosales et la luzerne (*Medicago truncatula* [Mtr]) du genre Fabales. Les dernières plantes utilisées font parties du groupe des Eurosids I tout comme notre plante modèle. Au sein du groupe des Eurosids I, le colza (*Brassica napus* [Bna]), la moutarde (*Brassica rapa* [Bra]) et la plante modèle (*Arabidopsis thaliana* [Ath]) font partie du genre Brassicales, mais leur famille est différente. Le colza et la moutarde appartiennent à la famille des Brassica alors qu'*Arabidopsis thaliana* fait parti de la famille des Arabidopsis.

Analyse générale des résultats de similitudes entre les séquences des organismes d'intérêt et les gènes d'*Arabidopsis*

En plus du critère d'éloignement vis-à-vis de la plante modèle, le nombre de séquences disponibles, autant EST qu'ARNm, est différent d'un organisme à l'autre. Ces deux critères ont donc été analysés, pour chaque organisme d'intérêt, afin de voir s'ils avaient un impact sur les résultats obtenus par Iccare. Le tableau 2.1 présente l'ensemble des résultats pour chaque organisme d'intérêt. La première colonne contient le nom des différents organismes utilisés. Les doubles colonnes 2, 3, 4 et 5 font références aux nombres de séquences EST et ARNm des organismes d'intérêt. La dernière colonne fait référence aux nombre de gènes similaires chez la plante modèle.

La première double colonne indique le nombre de séquences EST et ARNm disponibles dans les bases de données en janvier 2004 avant la récupération par Iccare. Les organismes d'intérêt peuvent être répartis en 3 groupes en fonction du nombre de séquences disponibles. Le groupe 1 est le groupe des organismes qui possèdent un grand nombre de séquences (plus

Organismes	**séquences dispo.**		**séquences util.**		**séquence simil. (e^{-20} <*E*-value≤ e^{-5})**		**séquences simil. (*E*-value< e^{-20})**		**Gènes modèles simil.**
	EST	ARNm	EST	ARNm	EST	ARNm	EST	ARNm	
B.napus	37159	968	36965	955	12126	342	19224	531	7486
B.rapa	6449	485	6360	483	2899	224	2808	207	3133
M.truncatula	198216	259	196290	258	70915	112	11841	21	8466
P.dulcis	3864	95	3857	95	890	18	128	9	675
H.annuus	59841	359	59590	356	18231	168	2176	46	3635
Z.mays	391158	15108	390071	14802	80706	4686	8190	608	5475
O.sativa	Ø	72561	Ø	72363	Ø	17185	Ø	1675	5680
P.patens	102226	490	100660	486	16170	166	852	41	1613
C.reinhardtii	154600	836	153932	818	8298	125	79	9	303

TAB. 2.1 – Résultat global de recherche de similitudes à partir des séquences EST et ARNm disponibles en janvier 2004.

de 100 000 séquences) dans lequel se retrouvent le maïs (*Z. maize*), la luzerne (*M. truncatula*), la mousse (*P. patens*) et l'algue verte (*C. reinhardtii*). Le riz (*O. sativa*), dont les séquences EST n'ont pas été récupérées, présente un grand nombre de séquences ARNm c'est pourquoi il est tout de même placé dans le groupe 1. A l'opposé, l'amandier (*P. dulcis*) et la moutarde (*B. rapa*) se retrouvent dans le groupe 3 (espèce qui possède un faible nombre de séquences, moins de 10 000). Le dernier groupe (groupe 2) représente les espèces qui possèdent un nombre de séquences intermédiaire (entre 10 000 et 100 000 séquences) et contient le colza (*B. napus*) et le tournesol (*H. annuus*). La deuxième double colonne contient le nombre de séquences EST et ARNm qui ont été utilisées par Iccare après traitement par *RepeatMasker* (élimination des séquences trop courtes, masquage des régions répétées et des vecteurs). Les troisième et quatrième doubles colonnes présentent le nombre de séquences EST et ARNm ayant des similitudes avec des gènes de la plante modèle avec respectivement une *E*-value comprise entre $1e^{-5}$ et $1e^{-20}$ et une *E*-value inférieure à $1e^{-20}$. Pour simplifier la lecture, j'utiliserai le terme *séquences de faibles similitudes* qui signifiera " séquences ayant des similitudes avec un gène de la plante modèle dont la *E*-value est comprise entre $1e^{-5}$ et $1e^{-20}$ ", le terme *séquences de fortes similitudes* qui signifiera " les séquences ayant des similitudes avec un gène de la plante modèle dont la *E*-value est inférieure à $1e^{-20}$ " et le terme *séquences similaires* qui signifiera " séquences ayant des similitudes avec un gène de la plante modèle dont la valeur est inférieur à $1e^{-5}$ " c'est-à-dire le cumul des *séquences de faible* et *de forte similitudes*.

Le nombre très variable de séquences disponibles avant analyse (variant de 10 000 à plus de 100 000 séquences) ne facilite pas la comparaison des résultats entre organismes. Afin de faciliter cette comparaison des données, les données de chaque double colonne ont été divisées par le nombre total de séquences disponibles (première double colonne). Ceci permet d'avoir accès aux ratios pour chacune des catégories et pour chaque organisme d'intérêt. Les résultats sont présentés dans la Figure 2.9.

Plusieurs résultats ressortent de ce graphique. Le premier est le nombre de séquences éliminées lors du traitement par RepeatMasker qui est très faible, au maximum 2% du nombre total de séquences, quel que soit l'organisme. Il y a donc très peu de séquence qui sont éliminées. Iccare peut travailler sur la quasi totalité des séquences récupérées. Cependant,

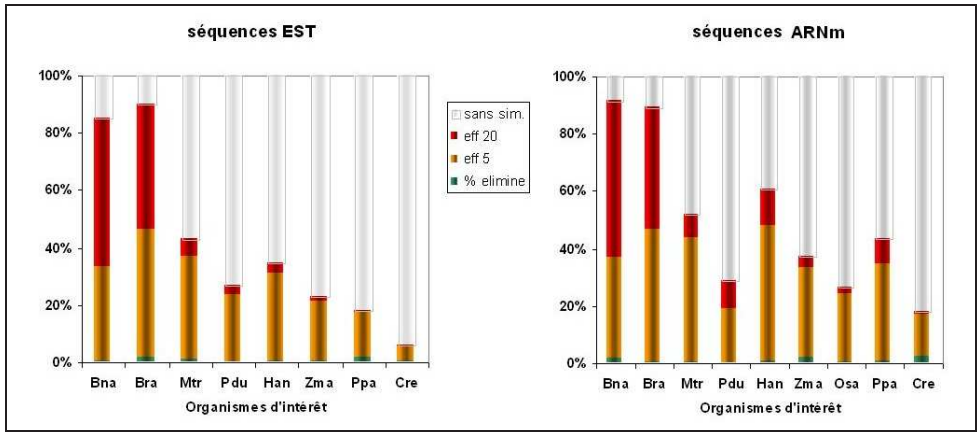


FIG. 2.9 – Représentation graphique du pourcentage des séquences EST et ARNm similaires à des séquences codantes de l'organisme modèle en fonction de l'organisme d'intérêt.

cette précaution du traitement par RepeatMasker évite d'avoir des problèmes par la suite lors de l'exploitation de ces séquences éliminées. Ensuite, le pourcentage des séquences qui présentent des similitudes (cumul des *séquences de faible* et *de forte similitudes*) décroît avec l'éloignement de la plante modèle. En effet, plus l'organisme d'intérêt est proche de la plante modèle plus le nombre de *séquences similaires* (*faible* et *forte similitudes*) est important, à l'exception du *Prunus*. Le colza [Bna] et la moutarde [Bra], qui sont deux plantes très proches taxonomiquement de la plante modèle, présentent plus de 85% de *séquences EST similaires*, alors que le maïs [Zma] et le tournesol [Han], un peu plus éloignées, ne présentent que 23 et 34% de *séquences EST similaires*. Pour les séquences ARNm, le pourcentage pour le colza et la moutarde augmente peu et passe à 90 %. Par contre, pour le maïs et le tournesol, les résultats augmentent et passent à 36 et 60 %. Ces résultats illustrent bien la diminution du nombre de *séquences similaires* avec l'éloignement de la plante modèle. Mais il est aussi intéressant de constater que les résultats sont meilleurs lorsqu'on utilise des séquences ARNm plutôt que les séquences EST. Ceci s'explique facilement par le fait que les séquences EST ne sont que des fragments partiels d'ARNm et que beaucoup d'entre elles correspondent à la partie UTR 3' ou 5' de l'ARNm (parties peu similaires entre organismes). Il est donc normal qu'un grand nombre de séquences EST ne présentent pas de similitudes comparés aux ARNm. En plus de la diminution du nombre de *séquences similaires* avec l'éloignement de la plante modèle, le nombre de *séquences de forte similitude* diminue aussi avec l'éloignement. A l'exception du *Prunus* [Pdu], tous les organismes présents suivent ce schéma. Le cas du *Prunus* est particulier car il présente un pourcentage de *séquences similaires* plus faible comparé aux autres plantes d'éloignement semblable [Mtr, Han]. Ces résultats peuvent s'expliquer par deux hypothèses : le génome de l'amandier est peu similaire à celui de la plante modèle ou le nombre de séquence utilisé pour la comparaison est trop faible pour avoir une vision globale de la conservation entre les deux espèces. Pour vérifier s'il s'agit de l'une ou de l'autre de ces hypothèses, il faudrait augmenter le nombre de séquences à comparer afin de voir si les résultats se rapprochent du taux de similitudes observé chez *Medicago* ou au tournesol qui sont d'éloignement semblable.

Après s'être intéressé à la répartition des séquences EST et ARNm en fonction de leur similitude, regardons comment se répartissent ces séquences vis-à-vis des gènes de la plante modèle. Plusieurs séquences EST et ARNm d'un organisme d'intérêt peuvent être similaires à un seul et même gène de la plante modèle. Pour faciliter la lecture, j'emploirai le terme *gène similaire aux ARNm ou aux EST* pour désigner " la séquence codante du gène de la plante modèle qui présente des similitudes avec des séquences EST ou ARNm de l'organisme d'intérêt ". Les gènes similaires ont été classés en fonction des organismes d'intérêt ainsi qu'en fonction du nombre de séquences EST ou ARNm avec lesquels ils sont similaires (1 gène similaire à : 1 seule séquence, entre 2 et 5 séquences, entre 6 et 10 séquences, entre 11 et 30 séquences et plus de 30 séquences). Les résultats de cette répartition sont présentés dans la Figure 2.10.

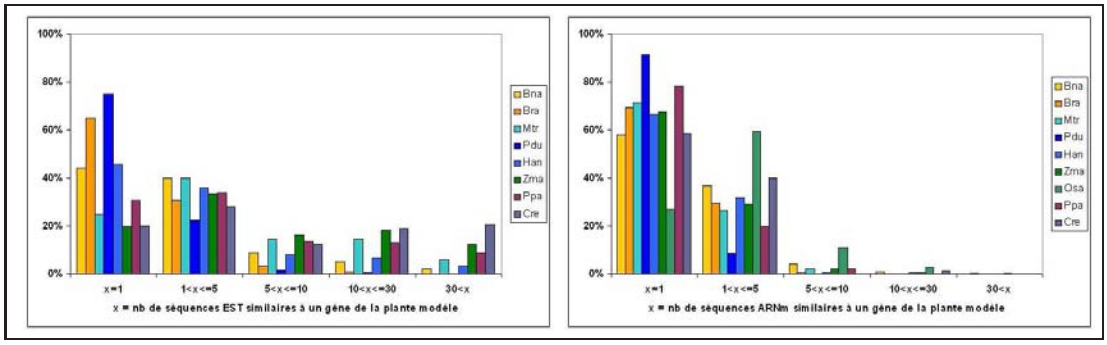


FIG. 2.10 – Répartition en pourcentage du nombre de séquences EST, à gauche, et ARNm, à droite, similaires à un gène de l'organisme modèle.

Les résultats présentés dans ces deux graphiques montrent la répartition du nombre de *séquences EST ou ARNm similaires* à un seul et même gène de la plante modèle. Cette répartition a aussi été étudiée en fonction de la qualité des similitudes (*séquences de faible similitude* et *séquences de forte similitude*). Les *séquences de faible* et *de forte similitudes* présentent les mêmes résultats (données non présentées). La qualité de la similitude n'influence donc pas la répartition des *séquences EST et ARNm similaires*. Concentrons-nous dans un premier temps sur les *séquences EST similaires*. Les organismes d'intérêt peuvent être regroupés en trois groupes. Le premier groupe est constitué de l'amandier [Pdu] et de la moutarde [Bra]. Dans ce groupe, la répartition est majoritairement de 1 séquence EST similiaire à 1 gène de la plante modèle. Le deuxième groupe est constitué du colza [Bna] et du tournesol [Han], dans lequel, la répartition est comprise entre 1 à 5 séquences EST similaires à 1 même gène de la plante modèle. Le troisième et dernier groupe (luzerne [Mtr], maïs [Zma], mousse [Ppa] et algue verte [Cre]) présente une répartition plus élevée d'environ 10 séquences EST similaires pour 1 même gène de la plante modèle. Cette répartition des résultats en 3 groupes correspond à la même répartition faite en fonction du nombre de séquences EST utilisées par Iccare. Le premier groupe contient la moutarde et l'amandier qui ont le moins de séquences EST et qui ont une répartition d'une séquence EST pour un gène de la plante modèle. De même, les espèces du deuxième groupe (tournesol, colza) possède respectivement environ 60 000 et 40 000 séquences EST et présente une répartition un peu plus élevée que

le premier groupe compris entre 1 et 5 séquences EST similaires pour un seul et même gène de la plante modèle. Et enfin le dernier groupe contient le reste des organismes qui possèdent plus de 100 000 séquences EST et une répartition d'environ 10 séquences EST pour un gène de la plante modèle. Il semble donc que plus le nombre de séquences EST disponibles est grand plus le nombre de séquences EST similaires pour un seul et même gène de la plante modèle augmente. Plus le nombre de séquences EST augmente plus les séquences EST sont redondantes.

Penchons-nous maintenant sur les *séquences ARNm similaires*. Les résultats pour chacun des organismes sont très similaires entre eux, à l'exception du riz [Osa]. Le nombre de séquences ARNm similaires est d'environ une séquence pour un gène de la plante modèle quel que soit l'organisme. Le nombre de séquences ARNm est plus faible que le nombre de séquences EST et les ARNm sont uniques alors que plusieurs séquences EST peuvent correspondre à un seul ARNm. Ceci explique la différence observée dans la répartition entre séquences ARNm et séquences EST. Les familles multigéniques peuvent en partie expliquer qu'un certain nombre de gènes de la plante modèle présentent des similitudes avec plusieurs ARNm (entre 2 et 5 séquences majoritairement). Les séquences ARNm des organismes d'intérêt ont évolué différemment des gènes de la plante modèle provoquant l'apparition de polymorphisme. Ces séquences, étant polymorphes, ne seront plus spécifiques d'un gène en particulier mais plutôt de la famille multigénique. Pour ce qui est du riz, les séquences ARNm disponibles très nombreuses (environ 72 000), on pourrait même dire trop nombreuses. Aujourd'hui, le nombre de gènes contenus dans le génome du riz est largement revu à la baisse (il est estimé à environ 50 000). Ce nombre surnuméraire de séquences ARNm a beaucoup été grossi par des séquences qui ont été annotées comme codantes pour un gène alors qu'il semblerait qu'elles soient le résultat d'événements de réinsertion dans le génome dûs à des transposons (Echenique *et al.*, 2002, Bennetzen *et al.*, 2004).

La dernière colonne du tableau présente le nombre de gènes de la plante modèle qui présentent des similitudes avec l'ensemble des séquences EST et ARNm (avec une *E*-value maximum de $1e^{-5}$). Pour faciliter la lecture, j'utiliserai le terme de *gènes similaires* pour désigner " les séquences codantes de la plante modèle qui présentent des similitudes avec des séquences EST et ARNm dont la *E*-value est inférieur à $1e^{-5}$ ". Les résultats présentés dans cette dernière colonne montrent qu'il y a de grandes variations dans le nombre de gènes similaires en fonction de l'organisme d'intérêt. Théoriquement, le nombre de gènes similaires devrait diminuer avec l'éloignement des espèces comparées. Pourtant les résultats observés ne sont pas concordants, par exemple *Arabidopsis* ne présente que 675 gènes similaires avec les séquences EST et ARNm de *Prunus* alors qu'avec *Medicago* il y en a 8466. En fait, pour analyser correctement ces résultas, il faut tenir compte de deux critères primordiaux : le nombre de séquences EST et ARNm utilisées pour la comparaison avec la plante modèle et l'éloignement vis-à-vis de celle-ci. Prenons l'exemple des deux plantes les plus proches d'*Arabidopsis*, le colza [Bna] et la moutarde [Bra]. Le colza possède environ 38 000 séquences EST et ARNm alors que la moutarde n'en possède que 7 000 par contre l'éloignement avec la plante modèle est sensiblement le même pour les deux organismes. Le nombre de gènes

similaires de la plante modèle est deux fois plus élevé avec le colza qu'avec la moutarde. Ceci montre bien l'influence du nombre de séquences EST et ARNm utilisées pour la comparaison sur les résultats. Plus ce nombre est élevé plus le nombre de gènes similaires est grand. Bien entendu, ce système est limité aux taux de conservation entre les deux organismes comparés. Même si le nombre de séquences EST et ARNm utilisées devient infini, le nombre de gènes similaires n'augmentera pas infiniment pour autant. Le nombre de gènes similaires est bien sûr limité par la conservation existante entre les gènes des deux espèces. Ceci est d'ailleurs bien illustré par l'exemple du riz et du maïs qui possèdent un grand nombre de séquences EST et ARNm, bien plus que le colza ou la luzerne, et qui pourtant ne présentent pas plus de gènes similaires que ces deux organismes. Pour ces deux plantes, on doit être proche du seuil de similitude entre le riz ou le maïs et *Arabidopsis thaliana*. Le deuxième critère est l'éloignement existant entre les organismes comparés. Le tournesol [Han] et le colza [Bna] ont approximativement le même nombre de séquences EST et ARNm (respectivement 60 000 et 38 000 séquences EST et ARNm), par contre leur éloignement avec la plante modèle est totalement différent. Le colza est très proche d'*Arabidopsis* alors que le tournesol est beaucoup plus éloigné. Le nombre de gènes similaires d'*arabidopsis* avec le colza est le double de celui observé avec le tournesol (respectivement 7 486 et 3 635). Le nombre de *gènes similaires* d'*Arabidopsis* avec le tournesol est le même que celui observé avec la moutarde qui possède pourtant environ dix fois moins de séquences EST et ARNm. Ceci montre bien que, pour un même nombre de séquences EST et ARNm, le nombre de gènes similaires est fonction de l'éloignement des organismes comparés.

Après nous être intéressés au nombre de *gènes similaires* de la plante modèle avec chacun des organismes d'intérêt, la répartition de ces *gènes similaires* a été étudiée en fonction de leur position sur les chromosomes de la plante modèle (Figure 2.11). Quel que soit l'organisme d'intérêt utilisé, cette répartition des *gènes similaires* est identique pour tous les chromosomes, à l'exception de l'amandier [Pdu] qui présente un nombre de *gènes similaires* un peu plus élevé sur le chromosome 1 et un peu moins élevé sur le chromosome 5. Mais cette différence de répartition est sûrement due au très faible nombre de séquence ARNm de l'amandier. Toujours est-il qu'il semble que les *gènes similaires* soient répartis de façon homogène dans le génome de la plante modèle.

Après avoir observé la répartition de ces *gènes similaires* au sein du génome de la plante modèle, j'ai regardé si ces gènes conservaient cette répartition en fonction du nombre d'organismes d'intérêt qui présentaient des similitudes avec le même gène. Les *gènes similaires* de la plante modèle ont donc été classés selon deux critères. Le premier concerne le nombre d'organismes d'intérêt qui présentent des similitudes avec chaque gène de la plante modèle. Neuf organismes d'intérêt ont été utilisés par Iccare, les *gènes similaires* de la plante modèle ont donc été classés en 9 groupes ; le premier contient tous les *gènes similaires* de la plante modèle qui ont des similitudes avec un seul des organismes d'intérêt (quel qu'il soit), le deuxième contient tous les *gènes similaires* avec uniquement deux organismes d'intérêt, le troisième contient les *gènes similaires* avec uniquement trois organismes d'intérêt, et ainsi de suite pour chaque groupe avec respectivement 4, 5, 6, 7 , 8 et 9 organismes d'intérêt. Le second critère

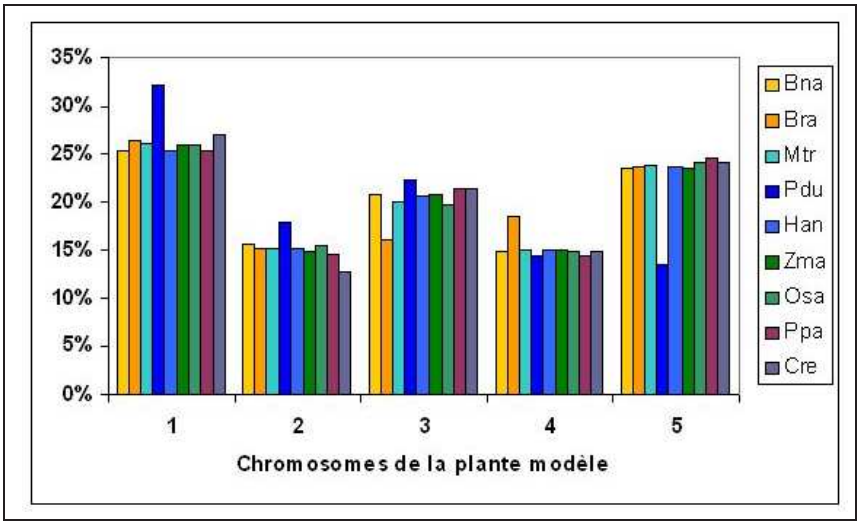


FIG. 2.11 – Répartition en pourcentage du nombre de gènes de la plante modèle en fonction des chromosomes.

étudié est la localisation sur les chromosomes de la plante modèle de ces *gènes similaires*. Cette répartition en fonction des chromosomes de la plante modèle est homogène pour chacun des groupes. Les *gènes similaires* sont donc répartis de façon homogène au sein du génome de la plante modèle (données non présentées). La répartition en fonction des groupes est présentée dans la Figure 2.12. Le graphique montre que la répartition des *gènes similaires* diminue avec le nombre d'organismes d'intérêt. Ces résultats ne sont pas surprenants dans la mesure où les organismes sont très éloignés les uns des autres, taxonomiquement parlant, et qu'il est donc normal qu'il y ait de moins en moins de gènes qui soient communs aux 9 organismes. Cependant ces résultats sont biaisés du fait que certains organismes utilisés dans notre étude ne possèdent qu'un petit nombre de séquences EST et ARNm. Le nombre de gènes similaires communs à 7, 8 ou 9 organismes est largement sous estimé. Mais il est tout de même intéressant de noter qu'il y a 2 343 gènes de la plante modèle qui présentent de similitudes avec au moins 4 organismes différents.

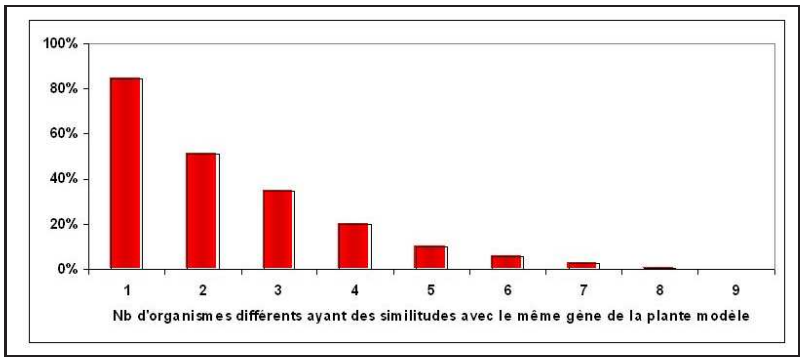


FIG. 2.12 – Répartition en pourcentage du nombre de gènes similaires de la plante modèle en fonction du nombre d'organismes d'intérêt.

Conclusions

Plusieurs conclusions peuvent ressortir de ces études. D'abord, les séquences ARNm sont plus informatives que les séquences EST. Les séquences ARNm présentent une meilleure proportion de *séquences similaires* avec un nombre moins important de séquences au départ. Cependant, les séquences ARNm disponibles ne représentent encore qu'une faible partie du génome exprimé d'un organisme (à l'exception du riz et du maïs). Pour pallier ce problème, les séquences EST peuvent être utilisées mais il faut un nombre important de séquences pour quelles soient informatives. Ensuite, il faut un nombre de séquences plus important avec l'éloignement entre l'organisme étudié et la plante modèle. Pour remédier à ce problème il faudrait, lorsque les informations seront accessibles, utiliser plusieurs plantes modèles qui serviraient de plante relais (ou plante nodale) à la plante modèle actuelle *Arabidopsis thaliana*. Ainsi, il suffirait de comparer l'organisme que l'on désire étudié non plus avec la plante modèle *Arabidopsis* mais avec la plante nodale la plus proche taxonomiquement. Les similitudes seraient ainsi plus fortes et le nombre de séquences informatives seraient plus important.

Les organismes qui présentaient le plus grand nombre de *gènes similaires* n'ont pourtant qu'un faible taux de conservation avec la plante modèle lorsque l'on compare le nombre de gènes similaires au nombre total de gènes dans le génome (environ 30 % pour le colza et la luzerne et environ 20 % pour le riz et le maïs). La différence de pourcentage entre colza/luzerne et riz/maïs peut s'expliquer par la différence d'éloignement entre les plantes et la plante modèle. De même, on peut se dire qu'entre le riz et le maïs, qui sont des monocotylédones, et la plante modèle *Arabidopsis*, la conservation ne doit pas être bien supérieure à 20 %, mais entre le colza et *Arabidopsis* ce poucentage est bien supérieur à 30 %. Il y a donc un biais dans ce calcul de pourcentage. Effectivement le pourcentage est sous-estimé et cela est dû d'une part à la méthode de traitement de Iccare et d'autre part à l'existence de duplications et de familles multigéniques au sein du génome d'un organisme. Les familles multigéniques existent chez tous les organismes et la plupart du temps ces familles multigéniques sont conservées, mais avec l'évolution les séquences de ces gènes divergent et les similitudes diminuent avec l'éloignement des organismes. L'ensemble des différentes séquences EST de l'organisme d'intérêt va donc présenter des similitudes non plus avec l'ensemble des gènes de la plante modèle mais avec certains des gènes de la famille multigénique. En plus de la divergence des séquences, Iccare associe les séquences EST à un seul gène de la plante modèle (celui qui présente la plus forte similitude). En fait, si ce gène appartient à une famille multigénique, la séquence EST ne sera associée qu'à un seul gène de cette famille multigénique. En fin de compte, alors que les séquences EST devraient être similaires à une famille multigénique (plusieurs gènes différents), Iccare introduit un biais en ne présentant que le meilleur résultat. Il faut donc être conscient que les résultats obtenus avec Iccare pour les différents organismes utilisés sont sous-estimés.

Un autre résultat intéressant montre qu'il existe des gènes de la plante modèle qui présentent des similitudes avec plusieurs organismes d'intérêt. Tous ces gènes peuvent ainsi servir à la construction de cartes génétiques comparées pour essayer d'établir la synténie existant entre ces espèces. De même, les *gènes similaires* à fort intérêt agronomique qui présentent

des similitudes avec plusieurs organismes d'intérêt ont de fortes chances d'être aussi retrouvés chez d'autres organismes qui pour l'instant ne présentent pas de séquences EST ou ARNm similaires. Ces séquences EST et ARNm n'ont pas encore été séquencées. L'intérêt de combiner les informations entre organismes permet d'estimer les gènes potentiellement similaires. De même avec les organismes pour lesquels peu de séquences EST ou ARNm sont disponibles. Ces organismes auront un nombre de *gènes similaires* faible mais les conservations d'autres gènes chez d'autres organismes peuvent être transposées à notre organisme.

En conclusion, Iccare est un outil efficace pour trier et analyser les grandes quantités de données en utilisant les informations de la plante modèle *Arabidopsis thaliana* comme référence. Même si la méthode d'analyse introduit un biais en ne sélectionnant que les séquences EST et ARNm qui présentent des similitudes, ces séquences sont fortement informatives. L'éloignement des organismes avec la plante modèle diminue le nombre de *séquences similaires* ainsi que de *gènes similaires*, mais cet obstacle peut être contourné en utilisant d'autres organismes modèles dès que les informations relatives à leur génome (et surtout aux séquences codants des gènes) seront disponibles (comme pour le riz, le peuplier et d'autres). L'incorporation d'un grand nombre d'organismes d'intérêt permet de transposer les informations des organismes ayant beaucoup de séquences EST et ARNm à ceux qui en ont beaucoup moins. La force de Iccare repose sur l'efficacité et la rapidité de l'analyse de grandes quantités de séquences et la possibilité combiner les informations de différents organismes.

2.2 Iccare par la Pratique : la RuBisCo

Les séquences EST qui codent pour la sous-unité S de la rubisco ont été testées sur Iccare afin de vérifier expérimentalement l'efficacité des amorces et des sondes définies en utilisant le transfert d'informations fournies par Iccare. Les amorces ont été définies afin de vérifier la conservation de position des introns sur la séquence génomique de plusieurs cultivars de tournesol et les sondes ont été définies afin de cribler une banque de clones BAC de tournesol afin de récupérer les clones BAC contenant les gènes de la sous-unité S de la rubisco.

2.2.1 La RuBisCo, une enzyme indispensable

La Ribulose-1,5-Bisphosphate Carboxylase/Oxygénase, communément appelée RuBisCo, est une enzyme indispensable à la vie sur Terre. Cette enzyme intervient dans le mécanisme de photosynthèse chez les végétaux. Elle permet la formation de matière organique à partir du carbone oxydé inorganique de l'air. La rubisco est l'enzyme centrale de la réaction d'oxydation de l'eau (H_2O) et de réduction du CO_2 pour fabriquer de la matière organique (par l'intermédiaire du cycle de Calvin et Benson). Poutant, cette enzyme est assez inefficace. Alors qu'une réaction enzymatique classique permet la synthèse d'une centaine de molécules par seconde, la rubisco ne peut fixer que trois dioxydes de carbone par seconde. Du fait de

cette faible activité, les cellules végétales compensent en produisant plus d'enzyme. Les chloroplastes sont remplis de rubisco (pouvant représenter la moitié des protéines). C'est pourquoi la rubisco est l'enzyme la plus abondante sur Terre.

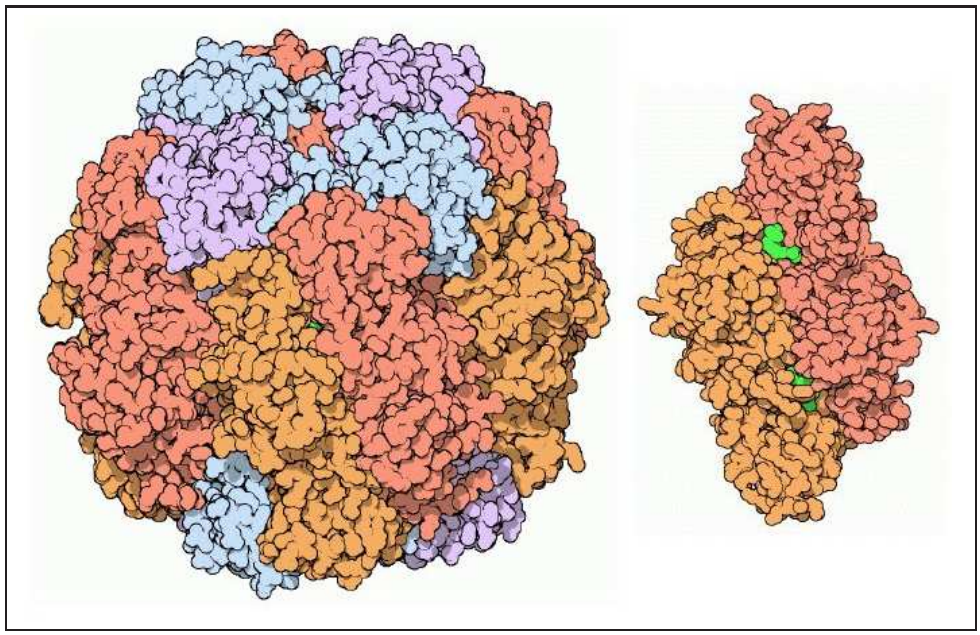


FIG. 2.13 – Représentation de la Ribulose-1,5-Bisphosphate CarboxylaseOxygénase d'après le site web "Protein Data Bank, molecule of the month (November 2000)".

La rubisco est aussi l'une des plus grosses enzymes de la nature avec un poids moléculaire de 560 kD (kilo Dalton). Chez les algues vertes et les plantes terrestres, le gène *rbcL* du chloroplaste code pour une sous-unité Large (Large subunit) de 55 kD, alors qu'une famille multigénique de gènes *rbcS* nucléaires code pour des sous-unités S (Small subunit) quasi identiques entre elles (15 kD). La rubisco existe sous deux formes : forme *I* et forme *II*. La forme *II* est l'association en dimère de sous-unités L, en orange et rouge à droite sur la Figure 2.13. Cette forme est retrouvée chez quelques procaryotes et des dinoflagels. La forme *I* est un hexadécamère constitué de 8 sous-unités L, en orange et rouge, et de 8 sous-unités S, en bleu et mauve, à gauche sur la Figure 2.13. Cette forme *I* est retrouvée chez les plantes terrestres, les algues vertes mais aussi certaines algues non vertes, bien que pour celles-ci les gènes *rbcS* soient chloroplastiques (pour plus d'informations lire la revue de Spreitzer et Salvucci, 2002). La fonction de la sous-unité S n'est pas encore vraiment connue. Il semble qu'elle influence la catalyse ainsi que la spécificité CO_2/O_2 (Esquivel *et al.*, 2002, Spreitzer, 2003).

Les gènes codants les sous-unités S de la rubisco

Les gènes *rbcS* de la rubisco ont beaucoup été étudiés à la fois pour essayer de déterminer leur fonction mais aussi pour mieux connaître cette famille multigénique. Les gènes *rbcS* font partie d'une famille multigénique dont le nombre de copies peut varier de 2 exemplaires à plus de vingt en fonction de l'organisme. Durant les 30 dernières années, plusieurs séquences d'ADNc et de gène *rbcS* de différents organismes végétaux ont été séquencées. Wolter *et al.* (1988) ont classé ces gènes *rbcS* en 4 catégories en fonction de la constitution en intron du gène *rbcS* (Figure 2.14).

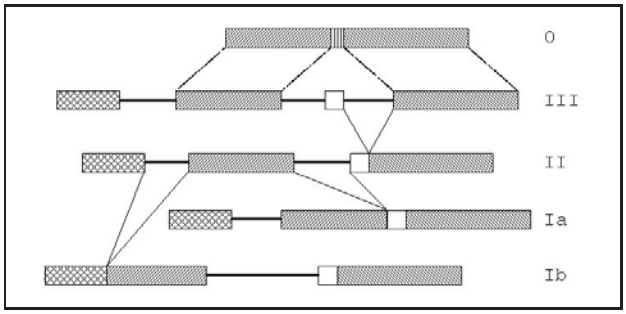


FIG. 2.14 – Structure des différents types de gène *rbcS* d'après Wolter *et al* (1988).

Le type 0 correspond aux gènes des procaryotes et correspond à des gènes sans intron (cyanobactéries). Les gènes du type III contiennent 3 introns (*Chlamydomonas reinhardtii*). Les gènes du type II contiennent deux introns et ont perdu le dernier intron (majorité des plantes terrestres et particulièrement les dicotylédones). Les gènes du type I contiennent un seul intron (les monocotylédones). Les gènes de type Ia possèdent uniquement le premier intron et correspondent aux plantes comme le blé (*Triticum aestivum*), le maïs (*Zea mays*). Les gènes de type Ib possèdent uniquement le deuxième intron et correspondent aux plantes comme la lentille d'eau (*Lemna gibba*). Le gène chloroplastique des procayotes possède deux régions (régions hachurées obliques) qui sont conservées dans les différentes formes du gène *rbcS* nucléaire. Mais les gènes nucléaires *rbcS* possèdent deux régions supplémentaires par rapport au gène procaryote. La première est localisée en début de séquence et correspond à un peptide d'adressage qui permet la migration de la protéine dans le chloroplaste. La deuxième séquence est venue s'intercaler entre les deux régions conservées. L'insertion de ces deux séquences est sûrement à l'origine de l'apparition des introns au sein du gène *rbcS*.

Les gènes *rbcS* de différents organismes végétaux

La forme ancestrale du gène *rbcS* (celle qui s'est retrouvée incorporée dans le génome nucléaire) semble être de type III (possède trois introns) comme retrouvé chez *Chlamydomonas*. *Chlamydomonas* possède deux gènes *rbcS* (quasi similaires) co-localisés dans son génome et contenant, tous deux, trois introns (Goldschmidt-Clermont et Rahiré, 1986). Au fil de

l'évolution, le troisième intron semble avoir été perdu pour former le type II. La petite plante appelé "Mésembryanthème à cristaux" (*Mesembryanthemum crystallinum*) possède 6 gènes localisés dans deux régions différentes de son génome. Tous ces gènes possèdent deux introns et sont plus ou moins similaires les uns aux autres (DeRocher *et al.*, 1993). La plante modèle *Arabidopsis thaliana* possède 4 gènes (1 gène sur le chromosome 1 et 3 gènes co-localisés en tandem sur le chromosome 5) qui ont chacun deux introns (Niwa *et al.*, 1997). Chez le pois (*Pisum sativum*), 5 gènes ont été caractérisés qui ont tous 2 introns et le haricot vert (*Phaseolus vulgaris*) possède au moins 3 gènes ayant 2 introns (Knight et Jenkins, 1992). Etrangement chez les Solanacées, il existe deux formes du gène *rbcS*, le type II et le type III. Des études menées sur la tomate (*Lycopersicon esculentum*, Sugita *et al.*, 1987 et Pichersky *et al.*, 1986) et la pomme de terre (*Solanum tuberosum*, Wolter *et al.*, 1988 et Fritz *et al.*, 1993) ont montré l'existence de 5 gènes *rbcS* localisés dans trois régions différentes du génome dont une possède trois gènes en tandem. Chez le pétunia (*Petunia hybrida*, Dean *et al.*, 1985) 8 gènes *rbcS* ont été trouvés dans trois régions différentes aussi dont une contenant 4 gènes. Ces trois plantes ont en commun d'avoir l'un de leurs gènes possédant 3 introns alors que tous les autres en ont deux. Les Solanacées semblent avoir conservé une version ancestrale du gène *rbcS* et cette forme est aussi retrouvée chez le tabac (Mazur et Chui, 1985). D'autres dicotylédones ont été étudiées depuis et le type II est toujours retrouvé (2 introns). Chez les monocotylédones, deux types de gènes *rbcS* ont été identifiés, le type Ia et Ib. Les gènes du type Ib possèdent un intron, qui correspond en fait au deuxième intron retrouvé chez les dicotylédones. La lentille d'eau fait partie de ces plantes, elle possède douze gènes *rbcS* qui ont tous un seul intron (Silverthorne *et al.*, 1990). Les gènes de type Ia possèdent un intron qui correspond au premier intron des dicotylédones. Le maïs, le blé, le riz et d'autres plantes font partie de ce groupe. Le nombre de gènes *rbcS*, chez ces individus, n'est pas vraiment identifié mais il semble que différentes classes de gènes *rbcS* puissent être établies au sein même de la forme de type Ia en fonction de la taille de l'intron (Sasanuma et Miyashita, 1998, Sasanuma, 2001).

Toujours est-il que la famille multigénique du gène *rbcS* est complexe et le nombre de gènes est différent d'un organisme à l'autre. De plus leur localisation n'est pas toujours constante. Par contre, la structure du gène est très bien conservée d'un organisme à l'autre. Le nombre d'introns est constant en fonction de l'organisme étudié (surtout chez les végétaux supérieurs tels que les dicotylédones dont les gènes ont 2 introns, hors Solanacées). De même, la structure de l'ARNm est toujours conservée avec un peptide d'adressage (une cinquantaine d'acides aminés en partie 5' de l'ARNm) puis la protéine mature. La comparaison entre séquences protéiques des gènes *rbcS* montre une bonne conservation entre les différents organismes.

Chez le tournesol

Le gène codant pour la grosse sous-unité (Large Chain, *rbcL*) est localisé dans le génome des chloroplastes, alors que ceux qui codent pour les petites sous-unités (Small Chain, *rbcS*) sont nucléaires. En revanche, le nombre de gènes codant cette petite sous-unité n'est pas

connu, toutefois un gène *rbcS* ainsi que l'ARNm correspondant ont été séquencés en 1987 (Waksman *et al.*, 1987, Waksman et Freyssinet, 1987). Ces deux séquences permettent de définir 2 introns pour ce gène *rbcS*, le premier d'une taille de 92 nucléotides et le deuxième d'une taille de 373 nucléotides. Depuis, le nombre de séquences EST issues d'ARNm de gènes *rbcS* a augmenté (actuellement environ 1 000 séquences), mais le nombre de gènes *rbcS* chez le tournesol n'est pas connu. Il est difficile de pouvoir définir combien de gènes *rbcS* codent pour ces sous-unités S. D'abord, parce que les séquences EST ont été obtenues à partir de différents cultivars de tournesol et aussi parce qu'un grand nombre de séquences EST est redondant (issues d'un même ARNm).

2.2.2 La Rubisco selon Iccare : les régions non codantes

Toutes les séquences EST et ARNm (récupérées en janvier 2004) ont été comparées aux séquences codantes de la plante modèle *Arabidopsis thaliana*. Cette analyse nous a permis, dans un premier temps de sélectionner toutes les séquences EST et ARNm qui présentaient des similitudes avec au moins un des 4 gènes *rbcS* codants pour la petite sous-unité de la rubisco chez la plante modèle (At1g67090, At5g38410, At5g38420 et At5g38430). Parmi les 59 590 séquences EST et les 356 séquences ARNm utilisées, 257 séquences EST et 2 séquences ARNm présentent des similitudes avec au moins un des gènes *rbcS* d'*Arabidopsis* avec une *E*-value inférieure à $1e^{-5}$. Vingt-cinq séquences EST ont été ajoutées à ces séquences (provenant du Génoplante). Toutes les séquences EST et ARNm ont été récupérées et soumises à la partie " Your Own Sequence Search " de Iccare afin de les comparer aux séquences codantes d'*A. thaliana* afin de comparer ces nouveaux résultats aux anciens. Les résultats sont présentés à cette adresse URL :http://genopole.toulouse.inra.fr/bioinfo/Iccare/pers_plante/vse_25May2005_15_10_48/.

Iccare ne permet pas d'assigner une séquence EST à son orthologue de la plante modèle, il permet juste de regrouper les séquences EST et ARNm qui présentent les plus fortes similitudes avec un des gènes *rbcS* de la plante modèle. Les séquences se répartissent donc en 4 groupes en fonction du gène *rbcS* d'*Arabidopsis* avec lequel les séquences présentent la plus forte similitude, à l'exception de 3 séquences (trop courtes ou similitudes trop faibles).

Les 4 groupes contiennent respectivement 16, 6, 6 et 253 séquences et sont respectivement assimilés aux gènes At5g38410, At5g38430, At1g67090 et At5g38420. Le groupe 1 assimilé au gène At5g38410 contient uniquement des séquences EST du laboratoire ayant de faibles similitudes (*E*-value supérieure à $1e^{-15}$). En revanche, tous les autres groupes présentent des séquences ayant de fortes similitudes (*E*-value inférieure à $1e^{-21}$). Iccare ne pouvant établir de relation d'orthologie entre les séquences de l'organisme d'intérêt et les gènes de la plante modèle, des études complémentaires sont nécessaires pour répondre à cette question. Par contre, Iccare permet de facilement déterminer l'emplacement des introns et des régions UTR (non traduites) sur les séquences EST et ARNm du tournesol à partir des information relatives aux 4 gènes d'*Arabidopsis*.

Chez les dicotylédones, la position des introns est très bien conservée, de plus on dispose de la séquence d'un gène *rbcS* et de son ARNm séquencé chez le tournesol (donc de la position des introns pour cette séquence). Les 4 gènes d'*Arabidopsis* ont leurs introns aux mêmes positions. La Figure 2.15 présente la comparaison de séquence entre le gène At5g38420 d'*Arabidopsis* et l'ARNm *X05079* du tournesol combiné à l'information de structure du gène d'*Arabidopsis* (découpage en exons). Chez *Arabidopsis* les gènes *rbcS* ont leurs introns respectivement entre les nucléotides 171 et 172 (il faut soustraire 21 nucléotides à la numérotation de la figure, celle-ci débute à la séquence EST et non au codon START du gène At5g38420, soit $192 - 21 = 171$ et $193 - 21 = 172$) et les nucléotides 306 et 307 (idem $327 - 21 = 306$ et $328 - 21 = 307$). La présentation graphique d'Iccare permet de facilement transférer la position des introns à nos séquences EST et ARNm. Ainsi, la séquence ARNm *X05079* possède une région UTR 5' non codante des nucléotides 1 à 21. Le codon START débute au nucléotide 22 et les deux introns sont situés entre les nucléotides 192-193 et 327-328. En revanche, la région UTR 3' n'est pas facile à établir dans la mesure où la similarité en fin de séquence est faible. L'outil de traduction, situé en bas au centre de la Figure 2.15, permet la comparaison des séquences nucléiques ainsi que la correspondance protéique de nos séquences. En débutant la traduction au nucléotide 22, la traduction des deux séquences permet de positionner le codon STOP (*TAA*) de la séquence ARNm du tournesol en position 556 à 558 (Figure 2.16).

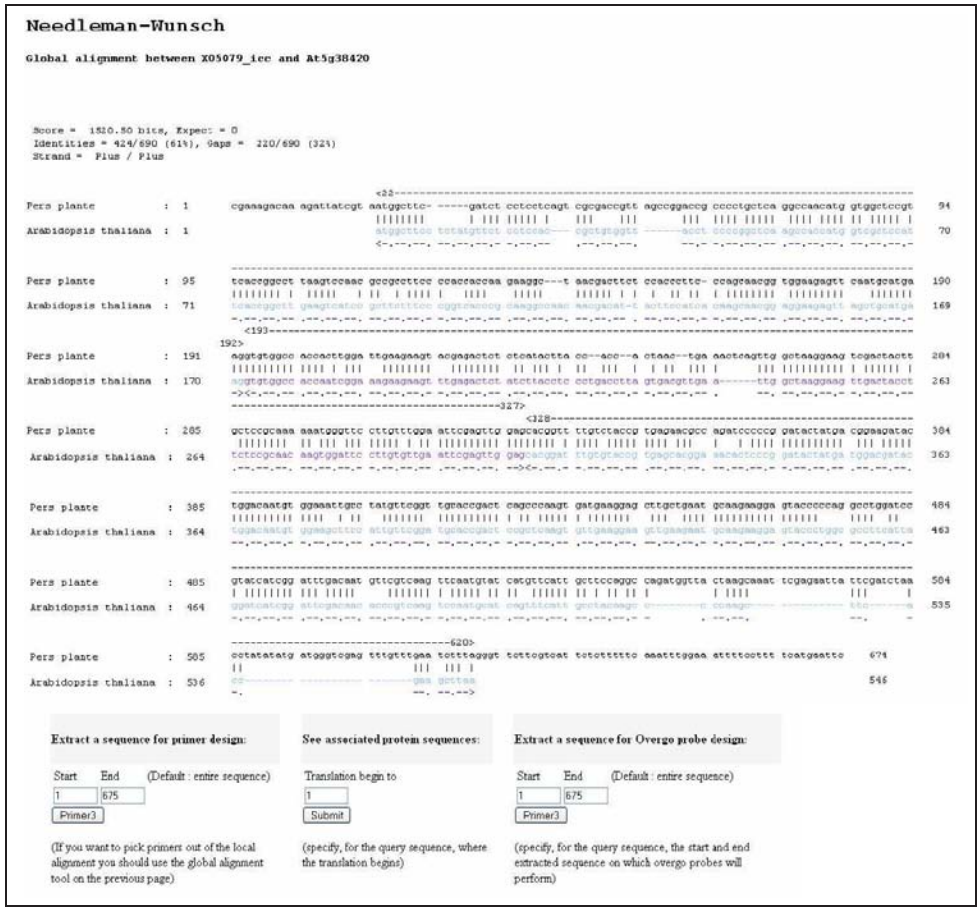


FIG. 2.15 – Position des introns sur le gène *rbcS* At5g38420 d'*Arabidopsis*.

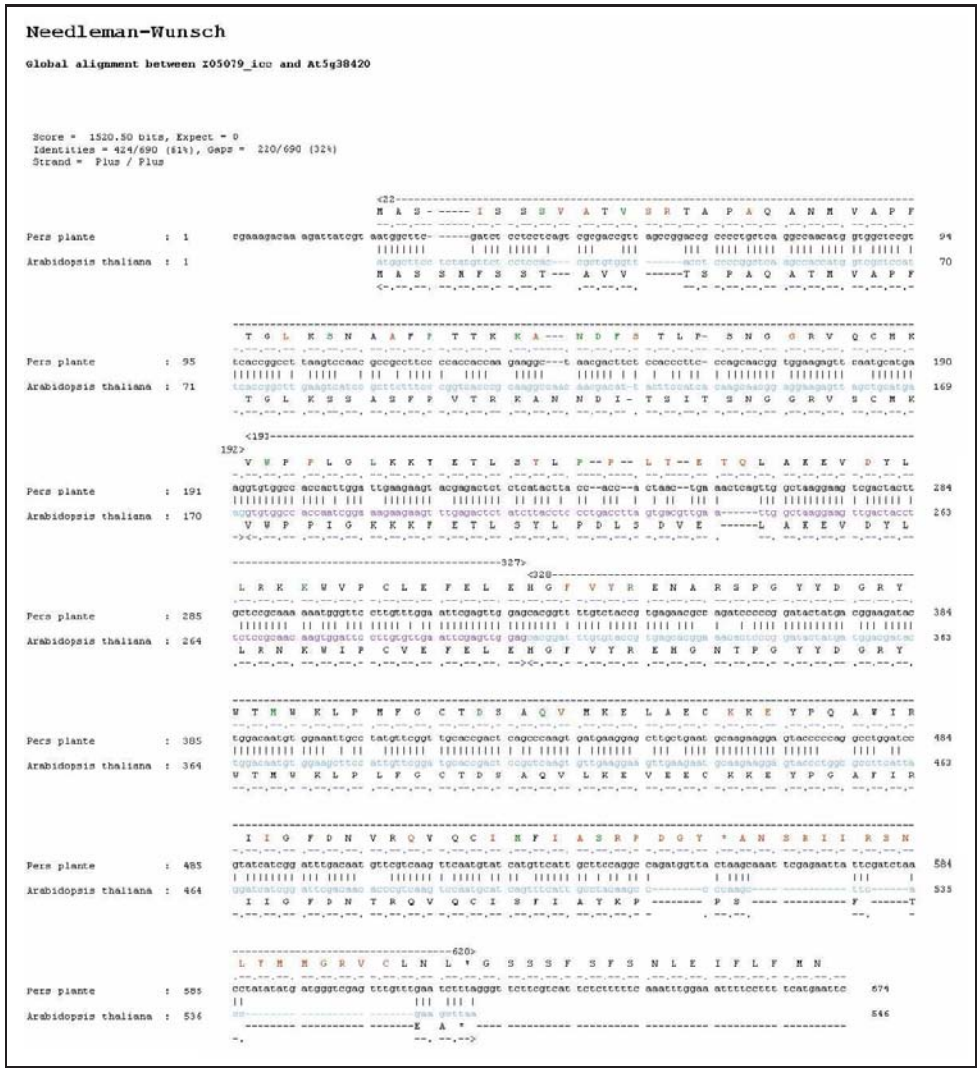


FIG. 2.16 – Comparaison d'une séquence EST et du gène *rbcS* At5g38420 d'*Arabidopsis*.

Pour faciliter l'analyse, le premier nucléotide du codon START sera numéroté 1. Le transfert d'informations du gène At5g38420 à notre séquence ARNm va nous permettre d'identifier les régions codantes des régions non codantes et la position des introns. En résumé, la séquence ARNm *X05079* est composée d'une partie non codante UTR 5' (nucléotides -21 à -1), du premier exon ($+1$ à $+171$), du premier intron (entre les nucléotides $+171$ et $+172$), du deuxième exon ($+172$ à $+306$), du deuxième intron (entre les nucléotides $+306$ et $+307$), du troisième exon ($+307$ à $+537$) et de la partie non codante UTR 3' ($+538$ à $+674$). L'alignement de la séquence ARNm avec la séquence génomique correspondante (GenBank ID : Y00431) confirme la structure prédite par Iccare (Figure 2.17), à l'exception de la position du premier intron. Mais ceci est uniquement dû au fait que le début de la séquence de l'intron1 et de l'exon2 ont 2 bases identiques "GT" (atgaag|GTtta...tcag|GTgtggc), c'est donc l'alignement qui a favorisé cette découpe alors que la région d'épissage est bien située entre les nucléotides $171 - 172$ (ag|GT) conformément à la loi d'épissage $GT...AG$. Le transfert d'informations structurelles d'un gène à un autre semble bien fonctionner.

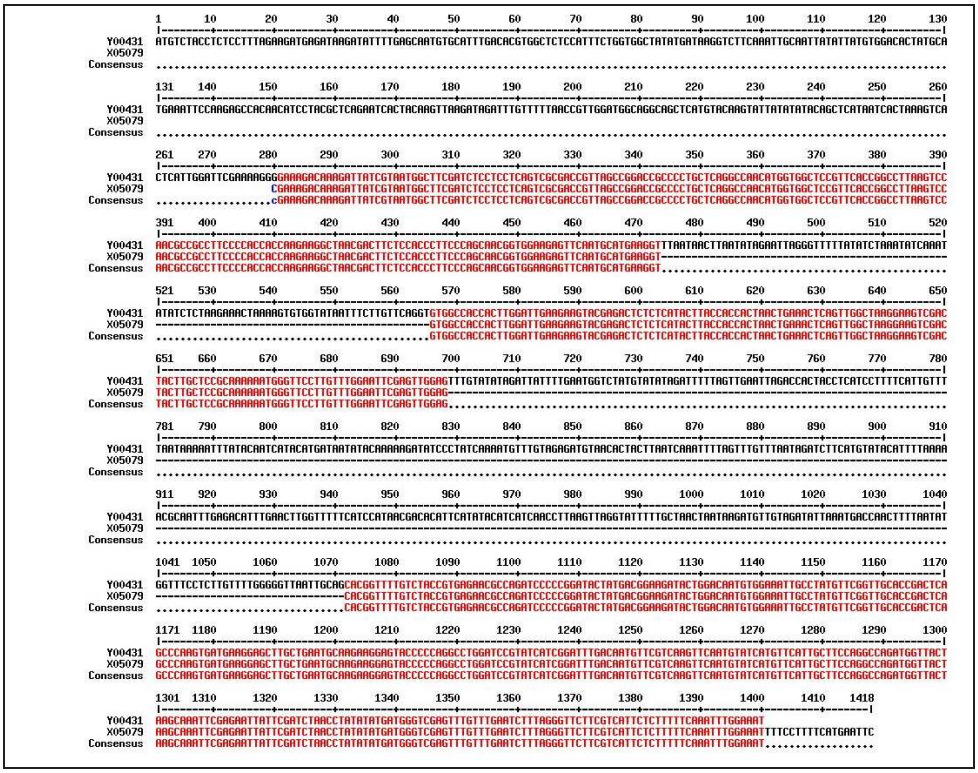


FIG. 2.17 – Alignement de la séquences ARNm avec la séquence génomique d'un gène *rbcS* du tournesol.

Bien que les séquences d'*Arabidopsis* et du tournesol présentent de nombreuses différences, celles-ci ne modifient pas le cadre ouvert de lecture ni la position des deux introns dans la séquence. Ces informations vont nous permettre de définir, dans un premier temps, des amorces de part et d'autre des introns pour vérifier expérimentalement s'ils sont bien présents, puis dans un deuxième temps, des sondes Overgo dans des parties exoniques conservées des séquences du gène *rbcS* pour cribler une banque de clones BAC.

2.2.3 Informations complémentaires

Avant janvier 2002, peu de séquences EST et ARNm de gène *rbcS* étaient disponibles pour le tournesol (la séquence ARNm publiée par Waksman et Freyssinet en 1987, des séquences EST du cultivar psc8 (une vingtaine dans la région 5' et une dizaine dans la région 3') provenant du GénoPlante). Un alignement dans la partie 5' de ces séquences a permis d'identifier des nucléotides polymorphes (SNP = Single Nucleotide Polymorphism = Nucléotide unique polymorphe) entre ces séquences (Figure 2.18).

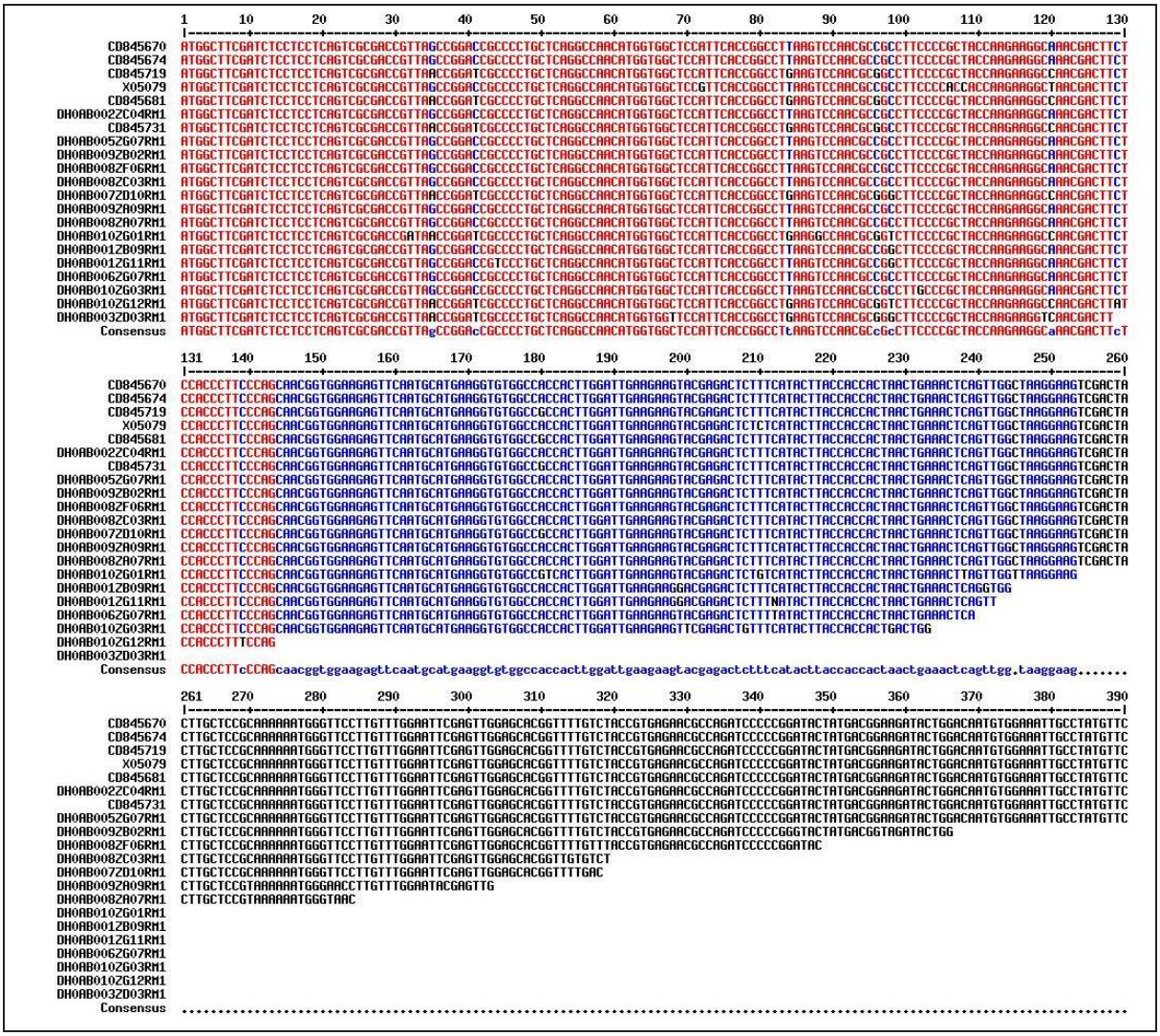


FIG. 2.18 – Alignement des séquences EST avec la séquence de l'ARNm *X05079* d'un gène *rbcS* du tournesol.

Pour déterminer les SNP de ces séquences, je considère qu'un nucléotide est polymorphe uniquement s'il apparaît dans plusieurs séquences. Si un nucléotide différent apparaît dans une seule séquence, ce nucléotide est considéré comme une erreur de séquençage, à l'exception

de ceux apparaissant dans la séquence de l'ARNm. Les SNP n'ont été déterminés qu'à partir du premier nucléotide du codon ATG, soit +1, jusqu'à la position du premier intron, soit +171. Trois principaux profils polymorphes se dégagent de cet alignement. Les trois profils ayant des SNP distincts présentent des nucléotides polymorphes aux positions +35, +41, +72, +84, +96, +98, +106, +108 et +120. Les différents profils polymorphes sont présentés dans le Tableau 2.2. Les nucléotides communs à plusieurs profils polymorphes sont mis en majuscules, et ceux minoritaires sont en minuscules.

Profils polymorphes	**Positions des nucléotides polymorphes**								
	+31	+41	+72	+84	+96	+98	+106	+108	+120
Profils ARNm	G	C	g	T	C	C	a	c	t
Profils psc8a	G	C	A	T	C	C/g/t	G	T	a
Profils psc8b	a	t	A	g	g	C/g/t	G	T	c

TAB. 2.2 – Différents profils polymorphes des séquences EST du cultivar psc8 et de l'ARNm *X05079* du gène *rbcS* du tournesol.

Les trois profils sont très proches les uns des autres. Cependant il reste un problème à résoudre au niveau du nucléotide +98 qui semble être polymorphe, mais le nombre de séquence ayant un G ou un T est bien plus faible que celles qui ont un C. Est-ce une erreur de séquençage ou un vrai SNP ? Les amorces qui ont été définies en combinant les informations de Iccare et ces informations de polymorphisme devraient pouvoir répondre à cette question.

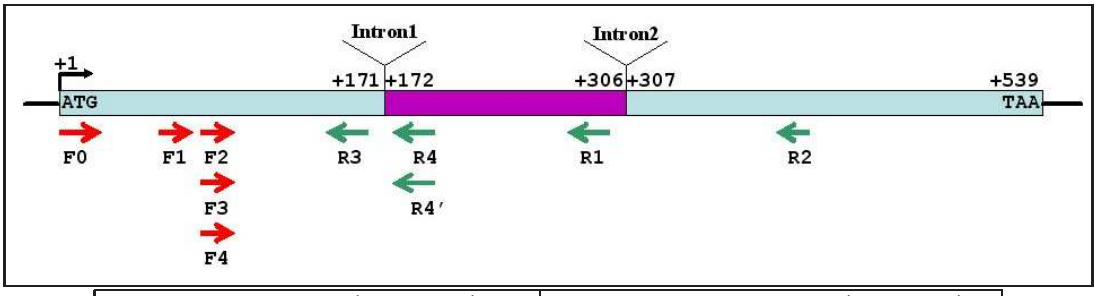
2.2.4 Vérification Expérimentale

Nous avons défini des amorces et des sondes Overgo à partir de la séquence ARNm et de séquences EST de tournesol qui présentaient des similitudes avec les gènes *rbcS* de la plante modèle *Arabidopsis*. Ces amorces vont nous permettre de vérifier qu'elles amplifient bien l'ADN génomique de tournesol, quel que soit le cultivar utilisé. De même, les sondes Overgo vont être testées quant à leur efficacité de criblage d'une banque de clones BAC.

Les amorces vont permettre de vérifier la présence des introns chez le tournesol et de séquencer le premier intron chez Ha300. Ces amorces serviront aussi à déterminer le nombre de gènes *rbcS* paralogues chez différents cultivars de tournesol en utilisant la technique de SSCP, Single Strand Conformation Polymorphism ou polymorphisme de conformation de l'ADN simple brin (voir l'annexe L page 188 pour plus de renseignements sur la technique de SSCP). Cette technique permet de différencier des fragments d'ADN ayant au moins un nucléotide substitué (un SNP au moins). Ainsi, nous devrions pouvoir identifier le nombre de gènes *rbcS* pour chaque cultivar mais aussi pouvoir différencier les gènes *rbcS* entre cultivars. Si les sondes Overgo fonctionnent, nous devrions pouvoir identifier dans la banque de clones BAC, l'ensemble des clones possèdant au moins un gène *rbcS*. Ces clones BAC pourrons alors être analysés pour déterminer l'organisation de la famille multigénique des gènes *rbcS* chez le cultivar Ha821.

Résultats

Définition des Couples d'amorces : les amorces ont été définies afin de pouvoir amplifier les régions introniques tout en tenant compte des polymorphismes repérés par l'alignement multiple. Toutes les amorces ont été définies par Primer3 à l'aide d'Iccare dans des régions fortement conservées à partir de la séquence ADN de l'ARNm *X05079*. Ces amorces sont présentées dans la Figure 2.19. Au total, 5 amorces sens (Forward) et 5 amorces anti-sens (Reverse) ont été définies. Pour tester les différents polymorphismes observés, 4 amorces ont été définies avec le dernier nucléotide discriminant (amorces F2, F3, F4 et R4). F2, F3 et F4 sont discriminants pour la base +98. F2 possède un *G* en +98, F3 possède un *C* et F4 possède un *A* (le *A* n'a jamais été observé en position +98, F4 servira donc de témoin négatif). L'amorce R4, quant à elle, est discriminante pour la base +180 et possède un *A*, elle s'oppose à l'amorce R4' qui possède un *N* (c'est en fait un mélange de 4 amorces ayant l'un des 4 nucléotides en +180).



Amorces Sens (Forward)		Amorces Anti-Sens (Reverse)	
Amorces	Séquences	Amorces	Séquences
F0	atggcttcgatctcctcctcagtc	R1	cccaaggaacaaaccttaagctca
F1	ggccaacatggtggctccat	R2	ccaacgtggctaagtcgggt
F2	ggccttaagtccaacgccgg	R3	gccaccttctcaagttacgtactt
F3	ggccttaagtccaacgccgc	R4	tggtgaacctaacttcttcatgct
F4	ggccttaagtccaacgccga	R4'	nggtgaacctaacttcttcatgct

FIG. 2.19 – Positions et séquences des différentes amorces définies à partir de la séquence ARNm *X05079* du gène *rbcS* du tournesol.

Amplification par PCR : des amplifications PCR sont réalisées pour les amorces F1, F2, F3 et F4 couplées aux amorces R1 et R2 ainsi que les amorces F0 et F1 couplées aux amorces R4 et R4' sur les 7 cultivars de tournesol (A : Ha300, B : psc8, C : KA, D : CP73, E : Ha821, F : PAC2 et G : RHA266). Le couple d'amorces de la calmoduline est utilisé comme témoin positif d'amplification sur 3 cultivars (A, B et C). Les résultats de la migration des amplifications sont présentés Figure 2.20.

Les témoins positifs d'amplification (calmoduline) ont fonctionné et une bande d'environ 1 400 pb (taille attendue) est observée pour les trois génotypes testés. Quel que soit le couple d'amorces utilisé, le témoin négatif (noté H) ne donne aucune bande prouvant l'absence de contamination lors des PCR. Le génotype PAC2 semble ne pas avoir fonctionné (pas de bande

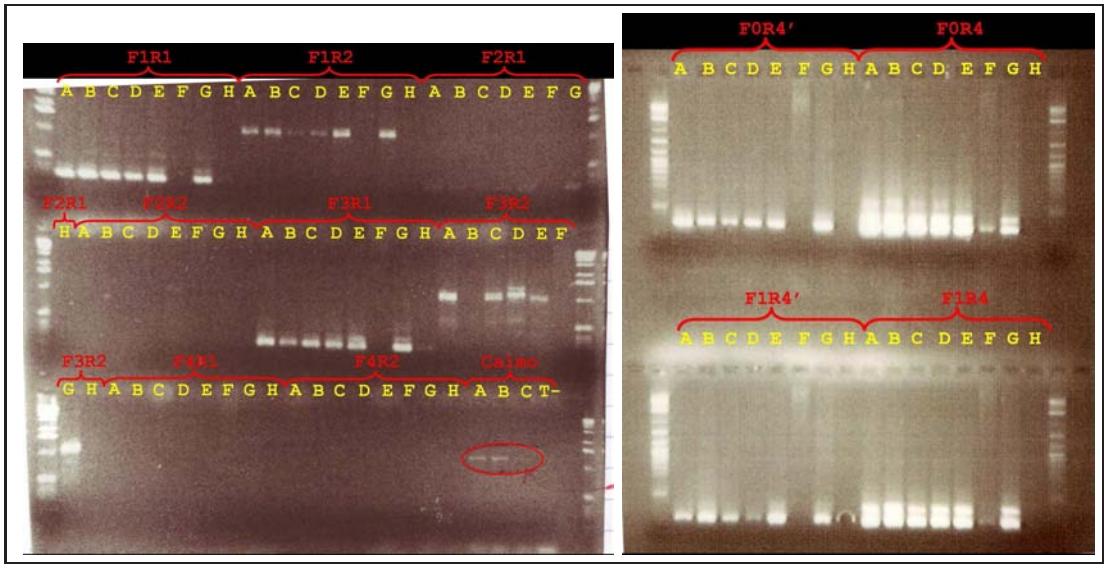


FIG. 2.20 – Migration des produits d'amplification sur gel d'agarose 2%.

visible), mais l'observation sous UV des gels montrait de faibles bandes non visibles. Ces résultats sont dûs à la mauvaise qualité de l'extraction de l'ADN du cultivar PAC2. Sur le gel de gauche, seuls les couples d'amorces F1R1, F1R2, F3R1 et F3R2 présentent une bande d'environ 350 pb (F1R1 et F3R1) et d'environ 1 000 pb (F1R2 et F3R2). Les autres couples d'amorces ne présentent pas de bandes. Sur le gel de droite, les produits d'amplification (F0R4 et F1R4) présentent une bande d'environ 250-300 pb alors que F0R4' et F1R4' présentent un amas de bandes plus épaisses.

Recherche de polymorphisme par SSCP : la migration en SSCP entraîne la formation sur le gel de deux régions distinctes en haut du gel et en bas du gel. Le haut du gel contient les fragments d'ADN simples brins, alors que le bas du gel contient les fragments ADN sous forme d'hétéroduplex. Seule la partie supérieure du gel nous intéresse. La migration des produits F1R1 et F3R1 (Figure 2.21) sur grand gel montre un profil de migration complexe. Le nombre de bandes par cultivar est difficile à établir, il semble varié de 10 à 16 bandes en fonction du cultivar. Cependant, ces profils de migration permettent de différencier les cultivars entre eux. Ainsi, trois profils de migration ressortent : un profil qui regroupe Ha300 [A], KA [C], Ha821 [E] et RHA266 [F], un deuxième profil contenant psc8 [B] et PAC2 [G] et le dernier profil contenant un seul individu CP73 [D]. L'amplification de F3R3 et F3R4 est de plus petite taille comparée à F1R1 et F3R1. La migration des produits d'amplification F3R3 est identique pour tous les cultivars et ne présente que deux bandes. Les produits d'amplification F3R4 présentent des profils de migration plus simples à distinguer. Le nombre de bandes visibles pour des produits F3R4 est compris entre 7 et 9 bandes en fonction du cultivar. De même qu'avec F1R1 et F3R1, les trois mêmes groupes de cultivars peuvent être établis en fonction du profil de migration.

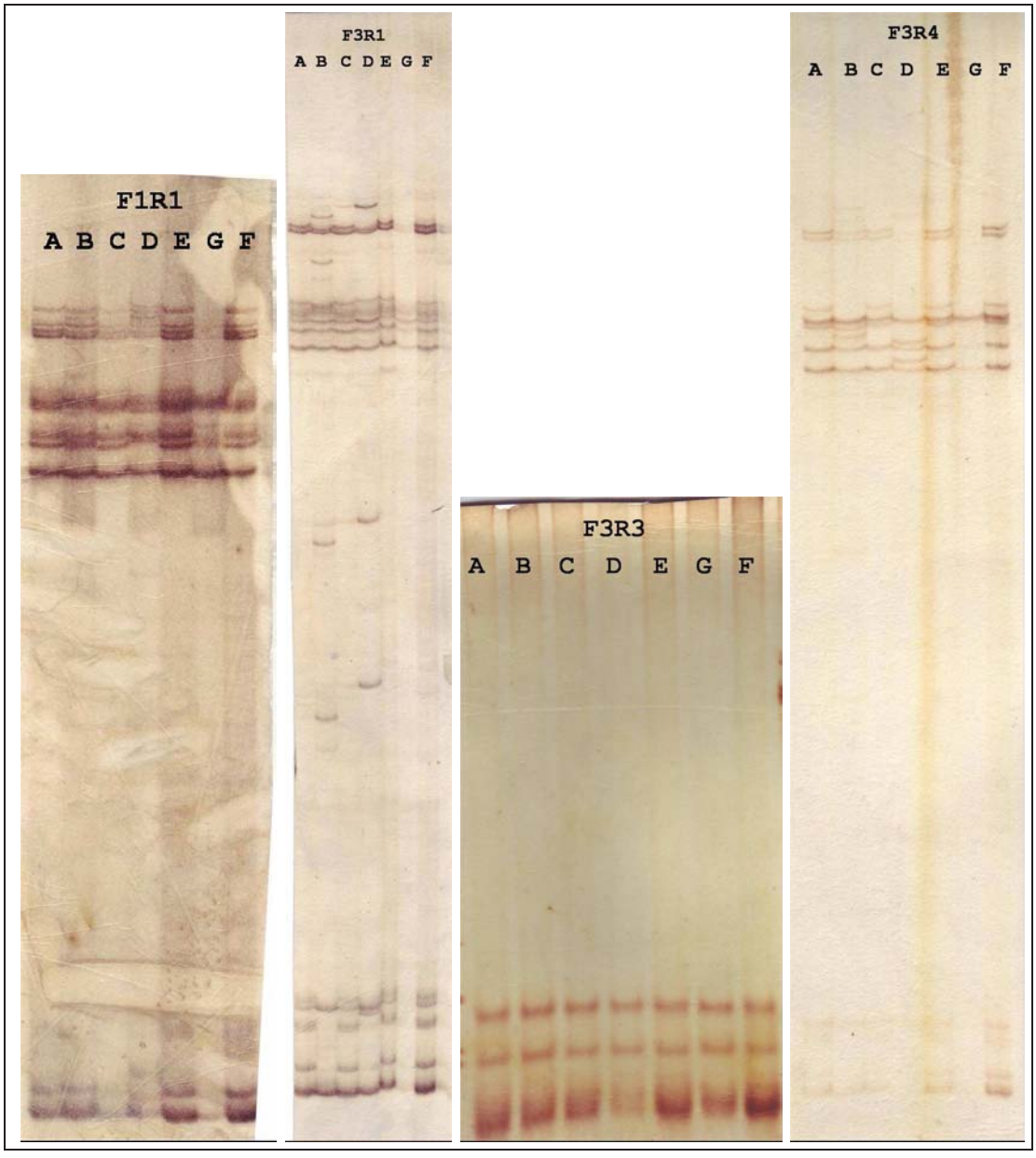


FIG. 2.21 – Migration des produits d'amplification sur gel SSCP.

Séquençage des produits d'amplification de Ha300 : les produits d'amplification de F3R1 et F0R4' ont été séquencés à partir du cultivar Ha300. Les séquences de F3R1 (24 séquences) et de F0R4' (28 séquences) ont été alignées pour déterminer le nombre de séquences polymorphes qui pouvaient être dinstinguées chez Ha300. F0R4' amplifie un fragment des positions +1 à +203 en amplifiant le premier intron et F3R1 amplifie un fragment des positions +79 à +301 y compris le premier intron. L'alignement de toutes les séquences montre que la séquence AIII19 est atypique comparée aux autres. Cette séquence présente un nombre plus élevé de nucléotides polymorphes, un codon supplémentaire dans le premier exon et surtout elle ne possède pas le premier intron. En dehors de cette séquence, toutes les autres peuvent être réparties en trois groupes en fonction des profils de polymorphisme avant l'intron. Même si les séquences F3R1 ne disposent pas du début de la séquence, elles peuvent tout de même être associées à celles de F0R4' d'après leur profil polymorphe. Le premier groupe nommé Ha300a est constitué de 11 séquences, le groupe Ha300b contient 19 séquences et le groupe Ha300c contient 22 séquences. Les profils de polymorphisme des séquences Ha300a, Ha300b et Ha300c sont identiques aux profils respectif de psc8a, de psc8b et de l'ARNm (Tableau 2.3).

Profils polymorphes	**Positions des nucléotides polymorphes**								
	+31	+41	+72	+84	+96	+98	+106	+108	+120
Profils ARNm	g	c	g	t	c	c	a	c	t
Profils psc8a	g	c	a	t	c	c	g	t	a
Profils psc8b	a	t	a	g	g	c	g	t	c
Profils Ha300a	g	c	a	t	c	c	g	t	a
Profils Ha300b	a	c	a	t	c	c	g	t	c
Profils Ha300c	g	c	g	t	c	c	a	c	t

TAB. 2.3 – Différents profils polymorphes des séquences EST du cultivar psc8, Ha300 et de l'ARNm *X05079* du gène *rbcS* du tournesol.

L'analyse des séquences avec l'intron permet de classer les séquences en 5 groupes. La Figure 2.22 présente quelques séquences de Ha300 représentatives des 5 différents profils polymorphes. En fait, il n'y a pas vraiment de nouveau profil polymorphe par rapport aux 3 profils déjà établis. La partie codante +1 à +171 ne présente que trois profils polymorphes représentés par les séquences AIII22, AIII26, AVII11 et AVII49 pour le profil Ha300a, les séquences AIII12, AIII13, AVII13, AVII39 et AVII12 pour le profil Ha300b et les séquences AIII30, AIII8, AIII28, AIII15, AVII8, AVII10, AVII21 et AVII47 pour le profil Ha300c. Par contre, il est clair que 5 profils différents sont présents dans l'intron, 2 associés à Ha300b, 2 autres à Ha300c et le dernier à Ha300a. Les introns du profil Ha300c sont très différents de ceux des profils Ha300a et Ha300b. Une seule base différencie les séquences des introns du profil Ha300c et deux bases différencient les séquences du profil Ha300b. Le profil Ha300c est polymorphe en position +27 du début de l'intron (+198 de l'ATG) et permet ainsi de classer les séquences Ha300c en 2 groupes. Le premier groupe noté Ha300c$\{G\}$ contient 12 séquences dont AIII30, AIII8, AVII21 et AVII47 ; le deuxième groupe noté Ha300c$\{T\}$ contient 10 séquences dont AIII28, AIII15, AVII8 et AVII10. Le profil Ha300b est polymorphe en positions +16 et +48 du début de l'intron (+187 et 219 de l'ATG) permettant de classer les

séquences en deux groupes. Le premier groupe noté Ha300b$\{GT\}$ contient 18 séquences dont AIII12, AIII13, AVII13 et AVII39 ; le deuxième groupe noté Ha300b$\{CC\}$ ne contient qu'une seule séquence AVII12. La séquence unique du profil Ha300b$\{CC\}$ est particulière car elle a un polymorphisme de type Ha300b avant l'intron mais l'intron correspond au polymorphisme de Ha300a. Au final, chez Ha300, le premier intron existe sous 5 profils polymorphes avec une taille de 91 ou 93 bases.

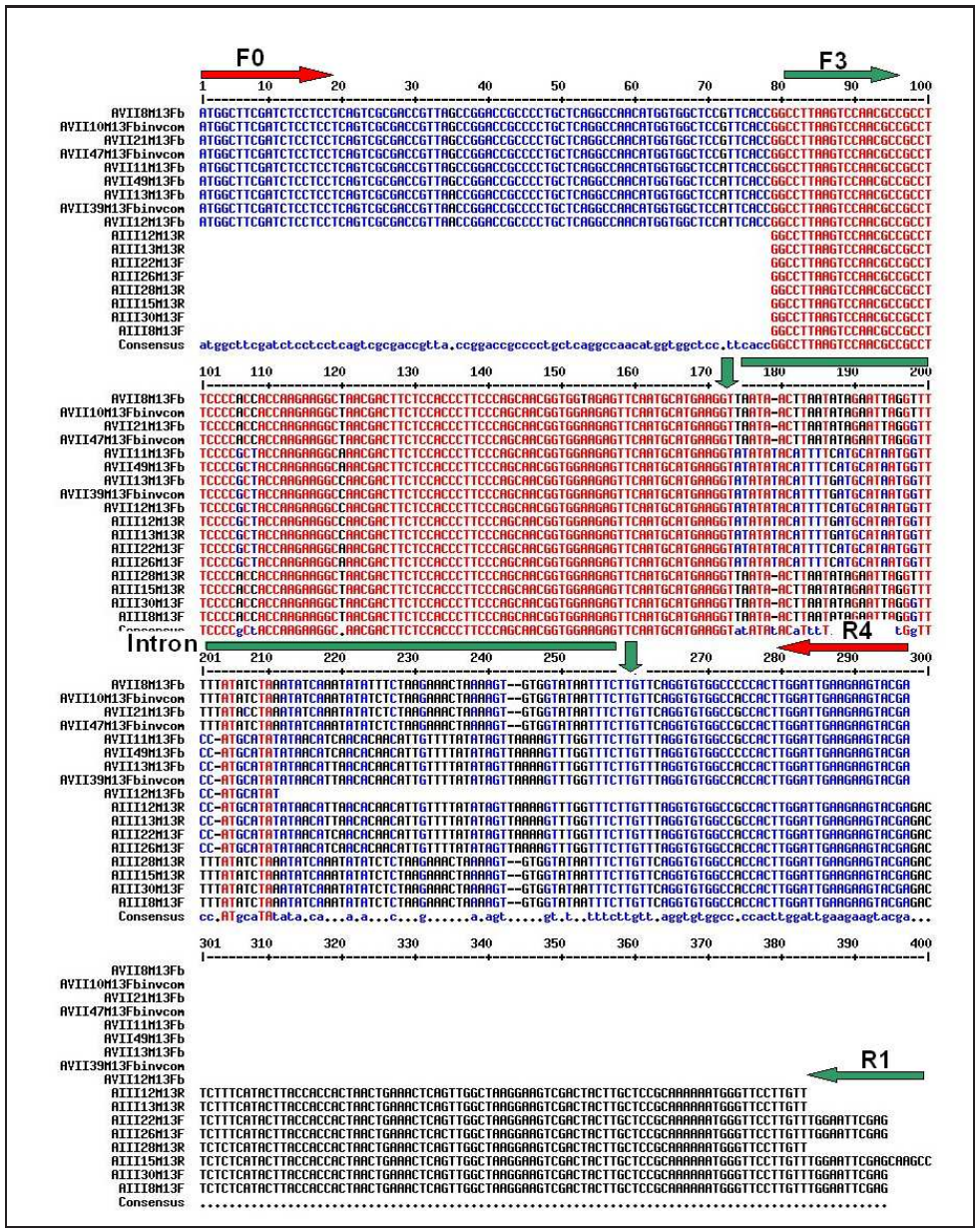


FIG. 2.22 – Alignement des fragments séquencés de Ha300 issus de l'amplification de F3R1 (AIII) et de F0R4' (AVII).

En ce qui concerne le nucléotide en position +180 (sans l'intron, sinon +273), les séquences issues de F0R4' (AVII) présentent un des 4 nucléotides avec une majorité de A quelque soit

le profil. Pour les séquences F3R1, le profil Ha300a et Ha300c ont un A en +180 alors que Ha300b a un G en +180.

Déterminer le nombre de gènes *rbcS* chez le tournesol : depuis Janvier 2004, un grand nombre de séquences EST similaires au gène *rbcS* sont disponibles dans les bases de données. Ces séquences ont été récupérées et regroupées en fonction du cultivar (Ha280, Ha801 et psc8). Seules 5 séquences sont disponibles pour le cultivar Ha801, le cultivar Ha280 dispose de 11 séquences et le cultivar psc8 en possède 273 (dont les vingt séquences déjà étudiées) ainsi que l'ARNm. En plus des séquences récupérées des bases de données, le séquençage à partir du cultivar Ha300 nous a permis de déterminer 5 profils polymorphes plus une séquence particulière, AIII19 (voir le paragraphe précédent). Les deux séquences ARNm dont nous disposons sont en fait identiques, par contre pour aucune des deux le cultivar n'est identifiable. Puisque les deux séquences sont identiques, seule la séquence ARNm *X05079* sera utilisée pour la comparaison.

Pour faciliter l'étude des nucléotides polymorphismes pour ces séquences, l'intron des séquences du cultivar Ha300 a été enlévé. Toutes les séquences ont été analysées uniquement pour les parties codantes (du premier nucléotide du codon START +1 jusqu'à la fin de la séquence). Les séquences de chaque cultivar ont ainsi été alignées avec MultAlin (Corpet, 1988, http://prodes.toulouse.inra.fr/multalin/multalin.html). Un nucléotide est considéré polymorphe uniquement s'il apparaît dans plusieurs séquences, à l'exception de ceux apparaissant dans les séquences ARNm, sinon ces nucléotides sont considérés comme une erreur de séquençage. Les profils de polymorphisme ont d'abord été établis pour chaque cultivar, puis les séquences les plus représentatives des profils polymorphes sont comparées entre cultivars.

Le trop faible nombre de séquences pour Ha801 ne permet pas de déterminer efficacement les différents types de polymorphisme. Pour Ha280, le problème est similaire, mais 7 séquences ont permis de déterminer 3 types différents de profils polymorphes (chaque profil contient au moins deux séquences). Les séquences disponibles pour psc8 sont beaucoup plus nombreuses et nous a permis, en alignant ces séquences, de retrouver les deux profils polymorphes déjà définis lors de nos analyses précédentes, page 67, ainsi qu'un nouveau profil qui est proche de celui de l'ARNm. Parmi les séquences un certain nombre d'entre elles sont des extrémités 3' et ne sont pas couvrantes jusqu'à l'*ATG*. Ces séquences n'ont été utilisées pour la recherche de polymorphisme (57 séquences). Les autres séquences (217 séquences) ont été regroupées en fonction du profil polymorphe (résultats non présentés). Nos trois profils ont donc été notés psc8a, psc8b et psc8c, et contiennent respectivement 54 séquences, 58 séquences et 105 séquences. Leurs profils respectifs sont présentés dans le tableau 2.4.

Cependant, il s'avère qu'en fait le profil psc8a présente du polymorphisme à la base +200. Le profil psc8a est composé en fait de deux profils polymorphes différenciés par une seule base. Ces profils sont notés psc8a$\{A\}$ et psc8a$\{G\}$ et contiennent respectivement 47 et 5 séquences (2 séquences du groupe n'ont pu être associées aux profils car elles finissaient avant la base polymorphe). De même, le profil psc8c peut être divisé en 2 groupes, mais le nombre de

Profils polymorphes	Positions des nucléotides polymorphes								
	+31	+41	+72	+84	+96	+98	+106	+108	+120
Profils ARNm	g	c	g	t	c	c	a	c	t
Profils psc8a	g	c	a	t	c	c	g	t	a
Profils psc8b	a	t	a	g	g	c	g	t	c
Profils psc8c	g	c	g	t	c	c	a	c	t

TAB. 2.4 – Différents profils polymorphes des séquences EST du cultivar psc8 et de l'ARNm *X05079* du gène *rbcS* du tournesol.

bases polymorphes qui les différencie est plus important qu'entre psc8a{A} et psc8a{G}. La première base qui les différencie est placée en position +237, les deux profils ont donc été notés, en fonction de cette base, psc8c{A} et psc8a{T}, et contiennent respectivement 43 et 68 séquences. Les résultats des séquences Ha300 sont présentés dans le paragraphe précédent.

Quelques exemplaires de chaque profil polymorphe pour nos 4 cultivars ainsi que la séquence de l'ARNm *X05079* ont été alignés pour rechercher les conservations de polymorphisme entre cultivars. L'alignement est présenté Figure 2.23. Les séquences commençant par BQ appartiennent à RHA801, celles qui commencent par BU sont de RHA280, celles en CD sont de psc8 et celles en AVII sont de Ha300. Des profils polymorphes sont retrouvés entre cultivars (à quelques modifications près). Ainsi, les profils polymorphes peuvent être regroupés en trois types principaux (Type a, Type b et Type c). Les profils polymorphes de psc8a{A}, psc8b et psc8c{T} sont respectivement utilisés comme profil référence pour chacun des types. Seul les nucléotides différents du profil référence sont indiqués dans le tableau 2.5.

Positions	+31	+41	+72	+84	+96	+98	+106	+108	+120	+180	+200	+210	+237	+308	+378	+401	+453	+456	+483	+542	+547	+548	+549	+550	+555	+562	+563	+564
Ha300a														Ø														
psc8a{A}	g	c	a	t	c	c	g	t	a	a	a	t	t	a	a	t	g	a	c	g	a	a	t	g	a	t	-	-
RHA280	Ø																											
psc8a{G}											g																	
Ha300b														Ø														
psc8b	a	c	a	t	c	c	g	t	c	g	a	t	t	a	a	t	g	a	c	a	g	g	a	a	t	a	-	-
RHA801		t		g	g										g		a				a							
Ha300c														Ø														
psc8c{T}	g	c	g	t	c	c	a	c	t	a	a	c	t	a	a	c	a	a	c	a	a	g	a	a	t	a	-	-
ARNm																	g	c	t									
RHA280																	g											
RHA801																	g											
psc8c{A}													a		g	t	g									t	a	
RHA280	Ø				a										g	t	g									t	a	

TAB. 2.5 – Profils polymorphes des séquences EST de 4 cultivars et de l'ARNm *X05079* du gène *rbcS* du tournesol.

Les 5 profils polymorphes de Ha300 sont alignés avec la séquence du gène *Y00431* du tournesol. La comparaison des séquences des différents introns montre que l'intron du gène *Y00431* est totalement identique à celui du profil Ha300c{T} à l'exception d'un nucléotide supplémentaire en +4 du début de l'intron (erreur de séquençage ou réel polymorphisme).

Définir des sondes : deux sondes Overgo sont définies dans les régions exoniques de la séquence ARNm *X05079*. La première est définie dans le deuxième exon en position +264

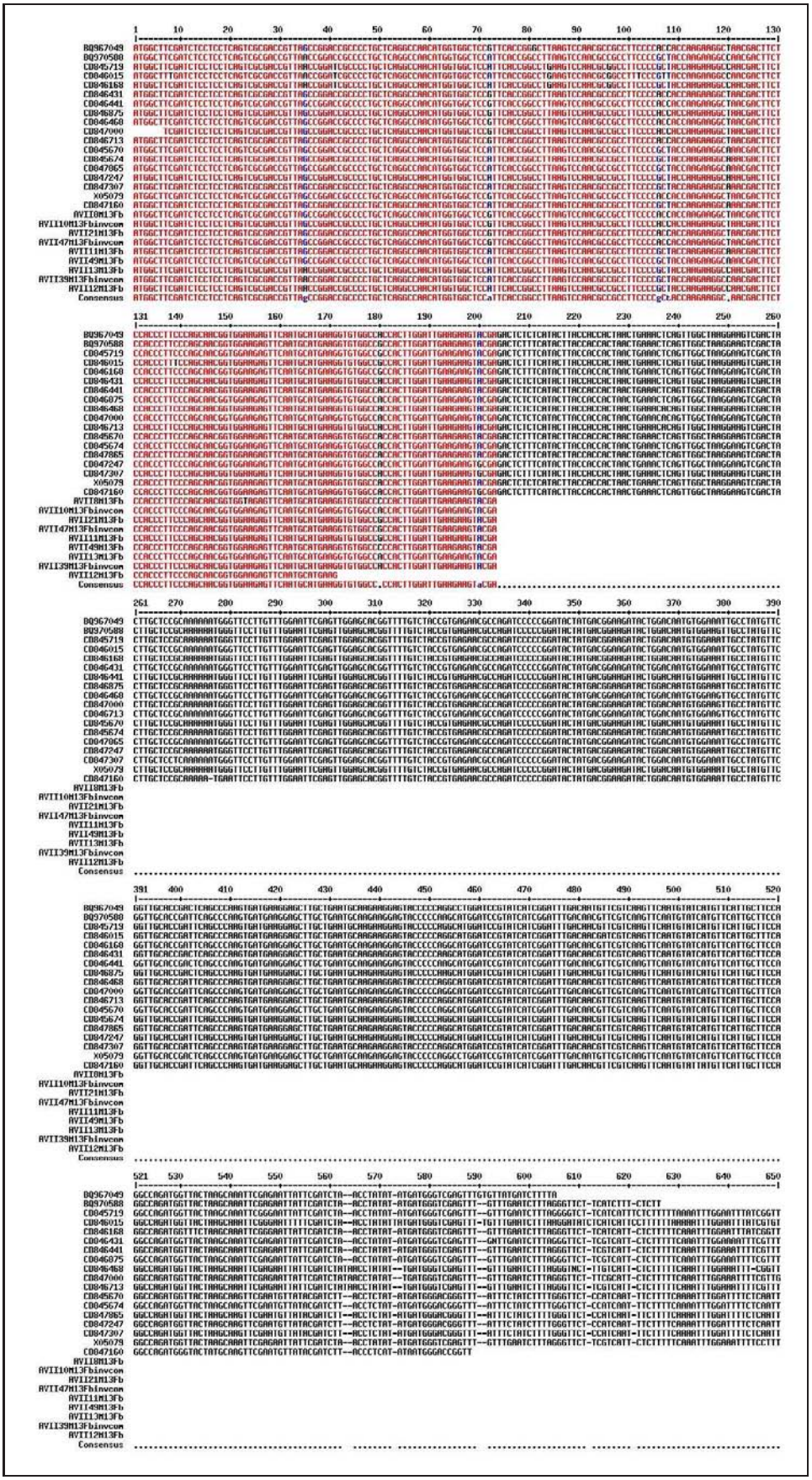


FIG. 2.23 – Alignement des différents profils polymorphiques de 4 cultivars différents de tournesol.

à +303 et la deuxième dans le troisième exon en position +437 à +476. Ces deux régions sont très bien conservées entre le tournesol et *Arabidopsis* et encore plus entre les différents profils polymorphes du gène *rbcS* pour le tournesol. Le pourcentage de *GC* de chaque sonde est différent. Rub1 a un pourcentage de *GC* de 45 alors que celui de Rub2 est de 50%. Les séquences des sondes Overgo Rub1 et Rub2 sont indiquées dans le Tableau 2.6. La partie en italique de la séquence est la zone de recouvrement de l'OVa avec l'OVb.

Sonde	Séquence OVa	Séquence OVb
Rub1	GCTCCGCAAAAAATGGGTTCCTTG	CAACTCGAATTCCAAACAAGGAAC
Rub2	TCAGCCCAAGTGATGAAGGAGCTT	CCTTCTTGCATTCAGCAAGCTCCT

TAB. 2.6 – Séquences des sondes Overgo Rub1 et Rub2 définies à partir de la séquence *X05079* du tournesol.

Criblage de la banque de clones BAC : le criblage de la banque de clones BAC a permis de récupérer un total de 20 clones BAC positifs avec les sondes Rub1 et Rub2. La sonde Rub1 n'a permis de récupérer que 5 clones BAC dont un seul n'est pas retrouvé par la sonde Rub2. Les 20 clones BAC positifs ont été extraits et des amplifications PCR avec les amorces F0R4 ont été faites sur ces clones. Le résultat de la migration est présenté Figure 2.24.

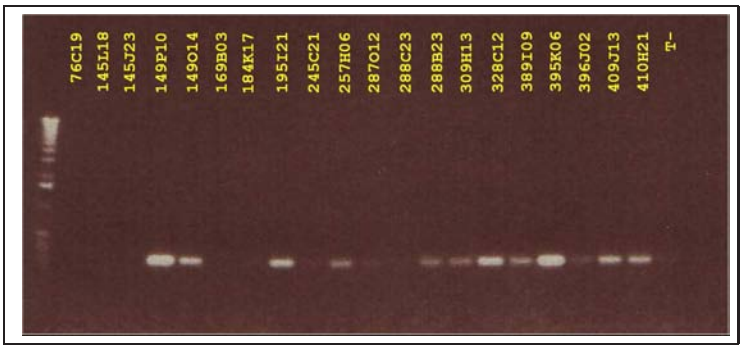


FIG. 2.24 – Amplification par F0R4 des clones BAC positifs avec les sondes Rub1 et Rub2.

Tous les clones BAC ne sont pas amplifiés. 7 clones BAC (149P10, 149O14, 195I21, 328C12, 395K06, 409J13 et 410H21) présentent une bande de forte intensité, 6 clones (257H06, 287O12, 288B23, 309H13, 389I09, 396I02) présentent une bande d'intensité plus faible et 7 clones n'ont pas été amplifiés (76C19, 145L18, 145J23, 169B03, 184K17, 245C21, et 288C23).

2.2.5 Discussions et Conclusions sur la Rubisco

Les amorces définies pour amplifier le ou les gènes *rbcS* du tournesol sont de deux natures différentes. Un premier lot composé de F0, F1, R1 et R2 a été défini dans des régions fortements conservées de l'ARNm *X05079*, fortements conservées avec le gène de *Arabidopsis* ainsi qu'avec les différents profils polymorphes du tournesol (à l'exception de F0 qui n'est pas

conservé chez *Arabidopsis*). Le deuxième lot est constitué des amorces F2, F3, F4, R4 et R4' et elles ont été définies de la même façon que pour le premier lot sauf que les amorces présentent une base discriminante en dernière position (côté 3'). Parmi les amorces définies, seules les amorces F2 et F3 n'ont rien amplifié. Pour les amorces F0, F1, R1, R4 et R4', l'amplification des gènes *rbcS* était attendue. Pour les amorces F2, F3 et F4, qui ne sont différentes que d'une seule base en position terminale (côté 3', position +98), seule l'amorce F3 a fonctionné, combinée à R1 elle amplifie deux fragments d'environ 300 pb et combinée à R2 elle amplifie un fragment d'environ 1 000 pb. L'amorce F4 (*A* en position +98, base qui n'avait jamais été observée sur les séquences de rubisco de tournesol) a été définie pour vérifier qu'il n'y a pas d'amplification, et c'est le cas. Les amorces F2 (*G* en +98) et F3 (*C* en +98) ont été définies pour tester le polymorphisme détecté par l'alignement des séquences EST (partie "informations complémentaires" page 67). Seule l'amorce F3 est amplifié, il semble donc que le polymorphisme observé lors de l'alignement des séquences résulte d'une erreur de séquençage. Le nucléotide en position +98 est donc un *C* chez le tournesol. Pour les amorces R4' et R4, elles ont été définies en position +203/+180 (ce sont des reverses, anti-sens) mais avec une différence de nucléotide en position +180. R4 possède un *T* (donc un *A* sur la séquence ARNm en position +180) alors que R4' est un mélange d'amorces ayant un nucléotide aléatoire en +180 (dernier nucléotide : *A*, *T*, *C* ou *G*). Un certain nombre de séquences présente un polymorphisme en position +180 avec l'une des quatre bases possibles. Les résultats d'amplification de ces amorces montrent que l'amplification F0R4' ne présente qu'une seule et unique bande alors qu'il y a un amas de bandes avec l'amorce R4. La non-spécificité de R4' semble permettre l'amplification des différents profils polymorphes (de tailles très semblable) grâce à la compétition engendrée par les différentes amorces, alors que R4, spécifique d'un profil polymorphe, doit entraîner l'amplification de profil hybride des gènes *rbcS* chez le tournesol ce qui explique l'amas de bandes sur le gel électrophorétique. Il semble donc que l'utilisation du dernier nucléotide comme base discriminante ne fonctionne que si le polymorphisme est de type présence/absence.

Revenons-en aux amorces qui ont fonctionné. Les amplifications F1R1, F1R2, F3R1 et F3R2 ont permis d'amplifier des fragments d'ADN visibles sur gel d'agarose. La migration des produits F1R1 montre deux bandes d'environ 350 pb avec une différence approximative de 10 pb, alors que la migration des produits F1R2 ne montre qu'une bande unique d'environ 1 000 pb. Avec des fragments d'environ 1 000 pb pour les produits d'amplification F1R2, s'il existe une différence de taille de 10 pb comme avec F1R1, elle ne sera pas visible sur ce type de gel d'agarose. Il est donc normal de ne voir qu'une seule bande pour les produits d'amplification F1R2. Les amplifiations F3R1 et F3R2 sont de tailles identiques à celles obtenues pour F1R1 et F1R2 car la différence entre les positions de F1 et de F3 n'est que de 25 pb. Aucun des produits d'amplification ne permet de discriminer les cultivars entre eux. En revanche, les produits d'amplification F1R1 (idem pour F3R1) confirment la présence du premier intron pour les différents cultivar du tournesol pour le gène *rbcS*. En effet, la taille des produits d'amplification F1R1 est supérieure à la taille d'amplification calculée s'il n'y avait pas d'introns (taille sans intron pour F1R1 : 248 nucléotides). Avec l'intron de la séquence du gène *rbcS* du tournesol publiée (*Y00431*), la taille des produits d'amplification doit être de 340 nucléotides (début

F1 +54, fin de R1 +301, Intron1 92 pb soit $(301 - 53) + 92 = 340$). La taille observée des fragments amplifiés par F1R1, environ 350 pb, confirme donc la présence d'un intron de taille similaire à celui du gène *rbcS* publié (92 pb). Pour vérifier la présence de cet intron, des produits d'amplification F3R1 et F0R4' ont été séquencés à partir du cultivar Ha300. Le premier intron a été retrouvé en même position que ceux des 4 gènes d'*Arabidopsis* et celui du gène *Y00431* du tournesol et présente une taille de 91 ou 93 pb en fonction du profil polymorphique de l'intron. Les produits d'amplification de F1R2 (idem pour F3R2) confirment la présence du deuxième intron pour les différents cultivars du tournesol. En effet, la taille des produits d'amplification F1R2 est supérieure à la taille d'amplification calculée en l'absence du deuxième intron (taille sans le deuxième intron pour F1R2 : 449 nucléotides). Avec le deuxième intron de la séquence du gène *rbcS* du tournesol publiée (*Y00431*), la taille des produits d'amplification doit être de 822 nucléotides (début F1 +54, fin R2 +410, Intron1 92 pb, Intron2 373 pb soit $(410 - 53) + 92 + 373 = 822$). Les produits d'amplification F1R2 ont une taille observée légèrement différente de celle calculée. Cette différence peut être due à la taille du deuxième intron (un peu plus grand) ou au niveau de résolution et au marqueur de taille utilisés qui sont moins adaptés pour cette taille de fragment. Toujours est-il qu'il y a bien deux introns dans les gènes *rbcS* du tournesol et quel que soit le cultivar utilisé. De plus, il n'y a pas de différence de taille pour les différents produits d'amplification F1R1 ou F1R2 entre les cultivars utilisés, ce qui implique que la taille des introns doit très peu variée d'un cultivar à l'autre.

Le nombre de gènes *rbcS* du tournesol ne peut être déterminé sur gel d'agarose. Les produits d'amplification F1R1, F3R1, F1R2 ou F3R2 ne présentent aucun polymorphisme de taille suffisante pour être visibles sur agarose. La migration des produits d'amplification a été faite sur des gels SSCP qui permettent une bien meilleure résolution que sur agarose afin d'essayer de déterminer le nombre de gènes *rbcS* polymorphes chez le tournesol pour différents cultivars. Les produits d'amplification F1R1 et F3R1 présentent un nombre de bandes visibles compris entre 10 et 16 en fonction du cultivar. Ces profils de migration sont complexe à analyser. Habituellement, la SSCP est utilisée pour distinguer deux gènes polymorphes entre eux et rarement des familles multigéniques. Dans notre cas le nombre de gènes et le polymorphisme existant entre eux ne permettent pas de distinguer le nombre exact de gènes codant pour la *rbcS*. Cependant, l'analyse des profils permet d'estimer le nombre minimum de gènes codant la *rbcS* à 5/8 gènes en fonction du cultivar. L'utilisation des produits d'amplification F3R3 (sans l'intron) montre le même profil de migration pour les différents cultivars avec deux bandes distinctes. Ces résultats semblent indiquer qu'il n'y a qu'une seule séquence amplifiée (ou plusieurs séquences de polymorphisme très comparables). En revanche, la migration des produits d'amplification de F3R4 (avec l'intron) permet de distinguer plusieurs bandes. On retrouve les trois mêmes groupes que pour F1R1 et F3R1 mais le nombre de bandes est plus facilement identifiable. Il varie de 7 à 9. En SSCP, le nombre de bandes ne peut être impair, s'il l'est c'est que deux fragments au moins ont co-migré. Le nombre de gènes *rbcS* chez le tournesol pour les amplification F3R4 doit être de 4 ou 5. Le premier intron des gènes *rbcS* de tournesol est donc une source de polymorphisme suffisante pour différencier plusieurs gènes. Mais il semble que d'autres nucléotides polymorphes se trouvent après l'intron (profil de migration F1R1 et

F3R1 plus complexe que ceux de F3R4). La SSCP n'a pas permis d'identifier le nombre de gènes *rbcS* existant chez le tournesol mais il s'agit bien d'une famille multigénique qui compte au moins 4 gènes *rbcS*.

Le séquençage des produits d'amplification F3R1 et F0R4' a permis d'identifier 5 profils polymorphes pour le cultivar Ha300 ainsi qu'une séquence (AIII19) qui présente un polymorphisme différent des autres séquences, un codon supplémentaire et surtout elle ne possède pas d'intron. L'absence de l'intron dans la séquence AIII19 semble suggérer qu'il s'agit d'une séquence ARNm qui aurait été intégrer au génome par des événements de tranposition (transposons ou rétro-transposons). En ce qui concerne les 5 profils polymorphes de Ha300, 2 des profils ne sont identifiables qu'avec la séquence de l'intron. Sans l'intron, seul trois profils polymorphes sont retrouvés. Ces trois profils comparés aux séquences des différents cultivars permettent de regrouper les séquences *rbcS* de tournesol en trois types polymorphes : le type a, type b et type c. Le type a regroupe les deux profils de Ha300a et de psc8a et celui de RHA280, le type b regroupe le profil de psc8b, de Ha300b et de RHA801b et le dernier type regroupe les deux profils de psc8c et de Ha300c, celui de RHA280 et de RHA801 ainsi que celui de l'ARNm (cultivar inconnu mais différent de ceux utilisés dans cette étude). Les cultivars ayant le plus de données sont psc8 et Ha300. Pour psc8, 3 profils différents dont 2 subdivisés en deux profils ont été identifiés, de même chez Ha300. On peut donc estimer le nombre de gènes *rbcS* pour ces deux cultivars à 5. Cependant chez Ha300, seule la séquence du premier intron a permis de distinguer deux profils pour Ha300a et Ha300c. Pour peu que psc8 possède aussi du polymorphisme au sein des introns qui permettrait de distinguer d'autres profils polymorphes en plus de ceux déjà détectés, le nombre de gènes *rbcS* pourrait doubler.

Il n'est pas évident de pouvoir déterminer le nombre exact de gènes *rbcS* existant pour un seul cultivar de tournesol. Seul le séquençage de tous les gènes et de leur localisation sur le génome permettra de répondre à la question. La localisation des gènes *rbcS* n'a pas encore été élucidée mais le criblage de la banque de clones BAC a permis d'isoler 20 clones BAC (dont 13 ont été confirmés par PCR). La sonde Rub1 n'a permis d'isoler que 5 clones BAC et Rub2 a permis d'en isoler 20, alors que toutes les deux ont été définies à partir de la même séquence (mais dans deux régions différentes). La sonde Rub1 a un pourcentage de G/C de 45% alors que celui de Rub2 est de 50%. La température d'hybridation est de 58°C pour la technique des sondes Overgo, cependant cette température est dépendante du pourcentage de G/C des sondes utilisées. Cette température doit normalement être légèrement diminuée lorsque le pourcentage de G/C est inférieur à 46%. La température utilisée pendant l'hybridation était de 58°C pour les deux sondes (car hybridées en même temps). La sonde Rub1 a donc été hybridée à une température un peu trop élevée, ce qui peut expliquer le plus faible nombre de clones BAC positifs. Par contre, les 5 clones BAC récupérés avec la sonde Rub1 sont très spécifiques. Les clones BAC récupérés devraient permettre d'identifier les gènes *rbcS* qu'ils contiennent et de déterminer l'ensemble des gènes de cette famille multigénique.

Les gènes *rbcS* font partie d'une petite famille multigénique. Iccare nous a permis de définir des amorces et des sondes qui se sont révélées très efficaces sur la rubisco. Les amorces ont per-

mis d'amplifier différents cultivars de tournesol et de séquencer un intron. Les sondes Overgo ont permis d'identifier des clones BAC contenant au moins un gène *rbcS* pour le cultivar Ha821. Bien entendu, d'autres études restent à mener sur ces gènes *rbcS* afin d'identifier tous les gènes *rbcS* ainsi que leur localisation. Il n'empêche que la méthodologie mise au point, basée sur le transfert des données d'*Arabidopsis* aux séquences du tournesol, fonctionne pour les gènes *rbcS* et devrait donc fonctionner pour n'importe quel autre gène. En effet, cette méthodologie nous permet d'optimiser l'exploitation des séquences EST et ARNm du tournesol grâce aux informations fournies par *Arabidopsis* afin de définir des amorces et des sondes qui soient efficaces pour l'analyse du génome du tournesol. Cependant, l'étude des gènes *rbcS* nous a aussi montré la difficulté de travailler à partir de gène appartenant à des familles multigéniques. Il est donc essentiel de mieux connaître les gènes à partir desquels seront exploitées les séquences EST de tournesol. Des informations complémentaires sur ces gènes nous seront donc utiles notamment pour interpréter les résultats expérimentaux (famille multigénique = risque d'avoir des profils d'amplification complexes, gène unique = une seule bande).

2.3 Exploiter la Synténie d'autres Organismes

2.3.1 Synteny Search, Rechercher la Synténie et les Duplications

L'étude de l'organisation et de l'évolution du génome d'*Arabidopsis* a fortement été aidée par la publication de la séquence complète de son génome (Mayer *et al.*, 1999, Lin *et al.*, 1999, Theologis *et al.*, 2000, Tabata *et al.*, 2000, Salanoubat *et al.*,2000). Il apparaît que plus de 75% des gènes d'*Arabidopsis* présentent des similitudes avec d'autres gènes d'*Arabidopsis* (Bancroft *et al.*, 2000). Le génome de cette plante a aussi subi de nombreuses duplications chromosomiques, reliques d'événements de polyploïdisation (Grant *et al.*, 2000, Vision *et al.*, 2000, Simillion *et al.*, 2002), et plus de 80% de son génome est dupliqué (Blanc *et al.*, 2000, Blanc *et al.*, 2004). Ces régions dupliquées ont subi des réarrangements et des événements de diploïdisation qui ont abouti à la perte de gènes. Cette évolution a mené à des gènes pour lesquels aucune similarité ne peut être détectée au sein du génome (Bowers *et al.*, 2003, Blanc *et al.*, 2004).

Aujourd'hui, l'utilisation du génome d'*Arabidopsis* pour faciliter les analyses de génomes d'autres plantes est bien plus difficile qu'il avait été espéré. Il est évident que connaître l'organisation et l'évolution du génome d'*Arabidopsis* est essentiel pour une meilleure exploitation de ces données dans les comparaisons de génomes. Il en est de même avec le génome du riz (*Oryza sativa*) qui est accessible depuis peu (Goff *et al.*, 2002, Yu *et al.*, 2002), même si les informations disponibles sont un peu moins nombreuses à l'heure acteulle.

Les familles multigéniques sont issues de ces évolutions des génomes et sont responsables des difficultés observées dans les études de comparaisons de génomes. L'utilisation des régions conservées entre la plante modèle *Arabidopsis thaliana* et des organismes d'intérêt nécessite qu'un gène chez la plante modèle corresponde à un seul gène chez l'organisme d'intérêt (gène

orthologue), sinon les résultats deviennent très rapidement difficilement analysables. Or, la très grande majorité des gènes appartient à des familles multigéniques ou sont dupliqués, ce qui rend difficile de déterminer quels sont les gènes qui sont réellement orthologues. Il est donc essentiel de savoir à partir de quel type de gène les informations sont transférées d'un organisme à l'autre. Un gène dupliqué ou appartenant à une famille multigénique chez la plante modèle a de grande chance de l'être aussi chez l'organisme d'intérêt et de ce fait de complexifier son analyse.

Afin de faciliter l'exploitation de l'ensemble des gènes du génome d'*Arabidopsis* ainsi que de ceux du génome du riz, le site web Synteny Search a été créé pour observer les relations des gènes entre eux au sein d'un même génome (duplications) ainsi que la synténie entre les gènes d'*Arabidopsis* et ceux du riz.

Synteny Search permet aux utilisateurs d'accéder à l'ensemble des gènes d'une région génomique et de visualiser les relations des gènes d'*Arabidopsis* ou du riz entre eux ou la synténie entre *Arabidopsis* et le riz pour n'importe quelle région du génome et de leur orthologue potentiel. Tous les gènes sont positionnés en fonction de leur localisation chromosomique et de leur orientation et l'utilisation de couleurs différentielles permet de discriminer les gènes uniques de ceux qui sont dupliqués. En plus, tous les gènes potentiellement paralogues ou orthologues peuvent être alignés et un arbre phylogénétique peut être construit à partir de ces données.

Programmation

Les relations entre gènes sont établis par la méthode utilisée par Paterson *et al.* (2000) et Blanc *et al.* (2004). Les séquences codantes ainsi que les informations de structure des gènes d'*Arabidopsis thaliana* sont récupérées sur le site web du MIPS (Schoof *et al.*, 2004) et celles du riz sont récupérées sur le site du TIGR, The Institut for Genomic Research. Les séquences d'un organisme sont ensuite comparées avec le programme BLASTn (Altschul *et al.*, 1997) à elles-mêmes pour la recherche de duplication puis à celles de l'autre organisme pour la recherche de synténie. Seules les séquences ayant des similitudes avec une *E*-value inférieur à $1e^{-5}$ sont considérées comme ayant une relation entre elles.

Les régions dupliquées ou synténiques sont définies lorsqu'au moins 25% des gènes dupliqués d'une région présentent des similitudes avec d'autres gènes localisés dans une même région chromosomique. Les résultats de cette analyse sont présentés sur un site web, Synteny Search.

Le site web

Le site web Synteny Search est hébergé par le Génopôle Toulouse Midi-Pyrénées. Il n'est pas encore totalement fonctionnel, mais on peut y accéder à l'adresse provisoire suivante URL :

http://genopole.toulouse.inra.fr/~cmuller/SynSearch/synsearch.html. Les pages intermédiaires contenant une synthèse des résultats sont à compléter, un certain nombre de liens sont encore à tester et la correction de "bugs" éventuels est encore à faire. Les pages qui permettent la visualisation du contenu en gènes des régions chromosomiques et des duplications sont fonctionnelles, en revanche, celles qui permettent de visualiser la synténie sont encore en cours de programmation.

La page d'accueil contient quelques informations sur le processus d'analyse du site, un lien vers le didacticiel "Tutorial" et 3 outils qui permettent de visualiser l'ensemble des gènes d'une région chromosomique, de visualiser les relations de duplications et de visualiser les relations de synténie entre *Arabidopsis* et le riz.

Le didacticiel permettra (il est en construction) de se familliariser avec le site web et les différents outils mis à disposition. La description du site ne sera pas faite dans ce manuscrit car n'étant pas fini, il est sujet à modifications. Cependant, la création des images contenant les résultats fonctionne (à l'exception de la synténie qui n'est pas encore achevée). Lorsque l'on définit deux régions à comparer ou que l'on observe le contenu d'une région en gènes, une image est créée en PNG de cette ou ces régions.

La recherche des duplications chez *Arabidopsis* est achevé et les résultats obtenus sont identiques à ceux obtenus par Paterson *et al.* (2000) et Blanc *et al.* (2004). Les résultats sont présentés dans la Figure 2.25 dans laquelle les différentes duplications du génome sont présentées. Chaque région est un lien vers la visualisation de la comparaison entre la région sélectionnée et celle dupliquée.

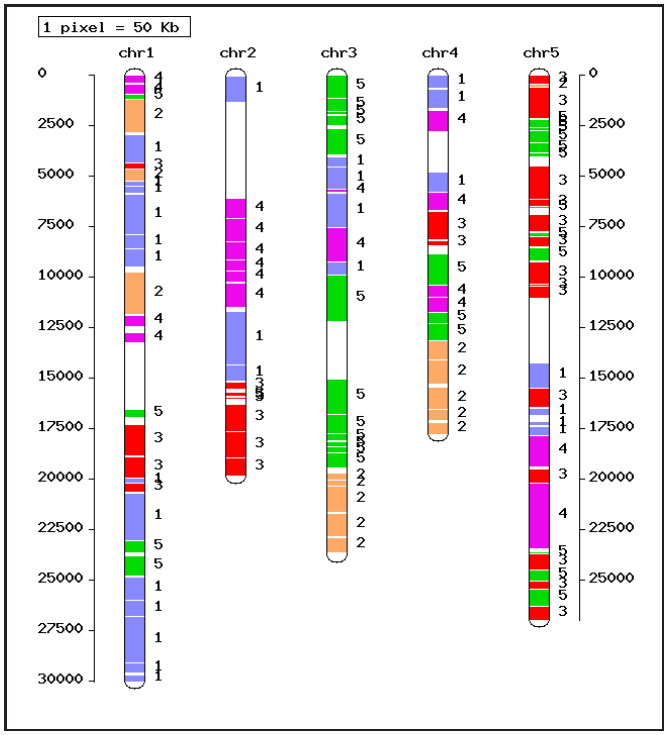


FIG. 2.25 – Duplication du génome d*Arabidopsis*, image créée par Synteny Search.

La Figure 2.26 est issue de l'observation du contenu en gènes et de la recherche de duplication et de synténie dans la région 4 250 à 4 750 du chromosome 5 d'*Arabidopsis*. L'image à gauche correspond à l'ensemble des gènes localisées dans la région 4 250 à 4 750 Kb du chromosome 5 d'*Arabidopsis*. Cette région contient 146 gènes et chacun des gènes est positionné en fonction de son orientation (strand + ou strand −). Le passage de la souris sur chacun des gènes permet d'ouvrir une boîte contenant des informations relatives au gène (position, sens, fonction potentielle, taille du gène, région cds). L'image du milieu correspond à l'ensemble des gènes de la région 4 25 à 4 750 du chromosome 5 d'*Arabidopsis* présentant des similitudes avec les gènes du riz (gènes en mauve et rose), alors que ceux en bleu ne présentent pas de similitudes avec des gènes du riz. Dans cette région, 63 gènes présentent des similitudes avec des gènes du riz. Par exemple, le gène At5g01390 présente des similitudes avec 4 gènes du riz dont 3 avec de fortes similitudes. De même sur la page "Content", lorsque la souris passe sur un gène une boîte de dialogue apparaît contenant le nom du gène ainsi que l'ensemble des gènes qui présentent des similitudes avec celui-ci. Cette image n'est pas la version définitive de la présentation de la synténie, il reste encore à assigner les gènes similaires aux représentations chromosomiques du riz. La dernière image à droite permet de visualiser les duplications qui existent chez *Arabidopsis*. Les régions 4 250 à 4 750 Kbases et 4 525 à 4 775 Kbases du chromosome 5 d'*Arabidopsis* sont dupliquée respectivement avec les régions 300 à 350 Kbases et 175 à 275 Kbases du chromosome 3. Tous les gènes de ces régions ne sont pas dupliqués. Les gènes ayant des paralogues potentiels sont classés en fonction de la localisation de leurs partenaires. Les couleurs utilisées sont le rouge, le vert, le jaune et le mauve et correspondent respectivement aux gènes qui présentent un partenaire localisé dans la région sélectionnée du chromosome dupliqué, localisé sur le même chromosome et dans la même région que le chromosome étudié, localisé sur le même chromosome que le chromosome étudié mais dans une autre région, ou localisé sur un autre chromosome (autre que celui étudié et dupliqué). Tous les gènes en rouge sont reliés à leur partenaire paralogue par un trait afin de visualiser la relation existant entre les deux régions. Une boîte identique à celle de la synténie apparaît lorsque l'on passe la souris sur un gène. Le nom du gène dans cette boîte de dialogue est un lien vers l'outil d'alignement multiple Multalin (Corpet, 1988 modifié, idem que pour Iccare). Cet outil permet de sélectionner quelques ou tous les gènes similaires et de les aligner. Les résultats de l'alignement peuvent être utilisés pour la construction d'un arbre phylogénétique (TreeTop, http://www.genebee.msu.su/services/phtree_reduced.html). Quelle que soit la page utilisée (Content, Duplication, Synteny), un tableau en bas de page permet de résumer l'ensemble des résultats : combien de gènes sont uniques, combien de gènes sont dupliqués.

En conclusion, Synteny Search permet, enfin permettra, de combiner les informations cartographiques et les relations entre les gènes afin de faciliter les études de comparaison de génome avec *Arabidopsis* ou le riz. L'addition de nouveaux organismes pourra se faire dès que les séquences complètes des génomes et leurs annotations seront accessibles.

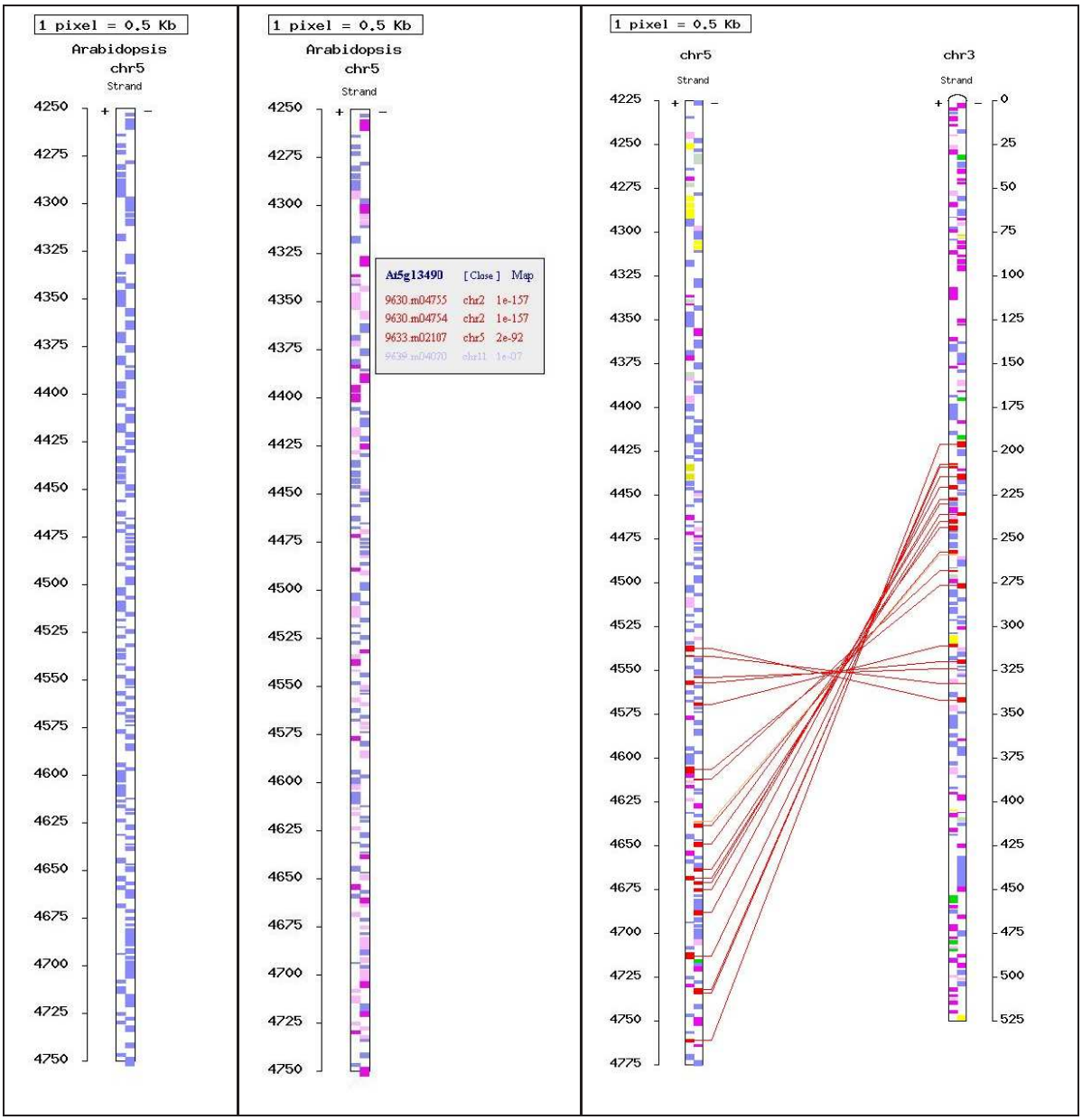


FIG. 2.26 – Image créée par Synteny Search avec le "Content" à gauche, la "Synteny" au milieu et la "Duplication" à droite.

Chapitre 3

Vérification Expérimentale

Le troisième objectif de mon travail de thèse est d'étudier la microsynténie avec *Arabidopsis*. Le premier objectif est atteint même s'il reste à confirmer l'efficacité de la technique à grande échelle. L'utilisation des sondes et d'amorces définies à partir de séquences EST ou ARNm de tournesol similaires à des gènes d'*Arabidopsis* va permettre d'analyser le génome du tournesol, d'émettre des hypothèses sur son organisation et sur la conservation de synténie avec celui d'*Arabidopsis*.

3.1 S'appuyer sur le Génome d'*Arabidopsis*

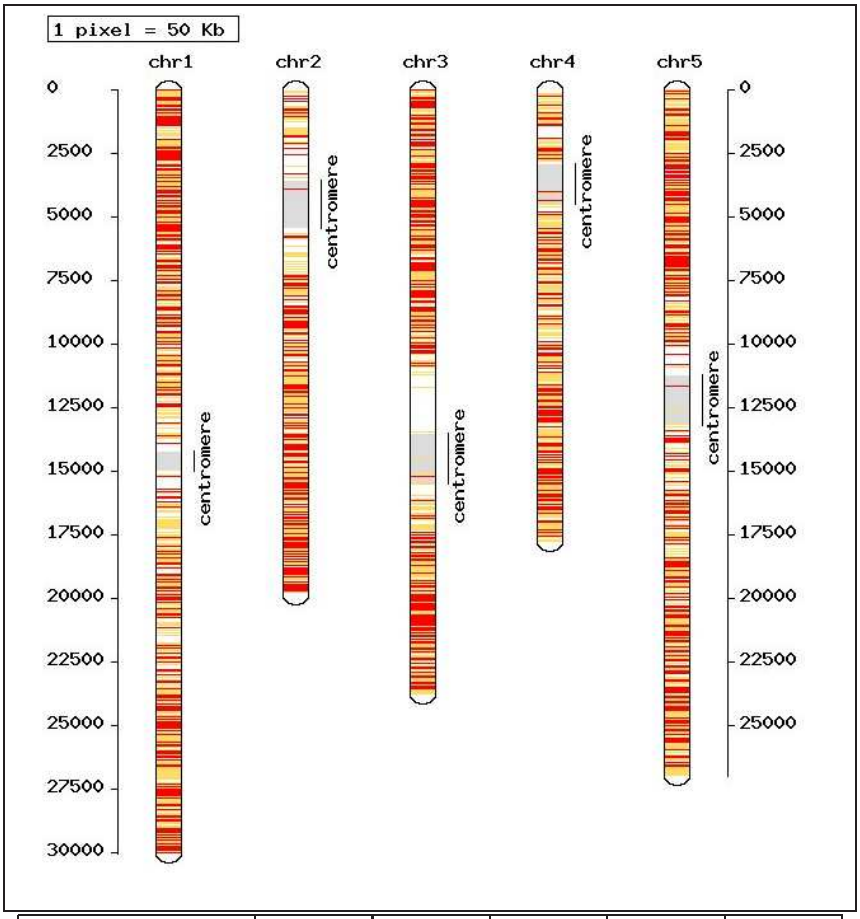
3.1.1 Transférer les informations du génome d'*Arabidopsis* à celui du Tournesol

Après la mise au point des outils bioinformatiques, nous avons défini un certain nombre d'amorces et de sondes Overgo d'abord pour vérifier l'efficacité de celles-ci à grande échelle, ensuite pour analyser l'organisation du génome du tournesol et de la synténie avec *Arabidopsis*. Nous avons donc recherché toutes les séquences EST et ARNm de tournesol similaires à des gènes d'*Arabidopsis* afin de définir des amorces et des sondes pour analyser le génome du tournesol.

Pour simplifier la lecture, j'utiliserai le terme *séquences de faibles similitudes* qui signifiera " séquences de tournesol ayant des similitudes avec un gène de la plante modèle dont la *E*-value est comprise entre $1e^{-5}$ et $1e^{-20}$ ", le terme *séquences de fortes similitudes* qui signifiera " les séquences de tournesol ayant des similitudes avec un gène de la plante modèle dont la *E*-value est inférieure à $1e^{-20}$ " et le terme *séquences similaires* qui signifiera " séquences de tournesol ayant des similitudes avec un gène de la plante modèle dont la valeur est inférieur à $1e^{-5}$ " c'est-à-dire le cumul des *séquences de faibles* et *de fortes similitudes*. De même, j'employerai

le terme *gène similaire* qui signifiera " gène de la plante modèle qui présente des similitudes avec des séquences EST et/ou ARNm du tournesol (inférieure à e^{-5}) ".

Un total de 60 200 séquences EST et ARNm a été traité par Iccare. Parmi toutes ces séquences, 18 231 séquences EST et 168 séquences ARNm présentent de *faibles similitudes* et 2 176 séquences EST et 46 séquences ARNm présentent de *fortes similitudes*. L'ensemble de ces séquences est similaire à 3 635 gènes de la plante modèle. Toutes les séquences EST ou ARNm sont potentiellement exploitables pour analyser l'organisation du génome du tournesol ainsi que la conservation de synténie avec *Arabidopsis*.



Similitudes	Chr. 1	Chr. 2	Chr. 3	Chr. 4	Chr. 5
$1e^{-20} > x$	3,9%	4,1%	4,0%	3,5%	4,1%
$1e^{-5} > x > 1e^{-20}$	9,4%	9,0%	9,8%	9,9%	9,8%
$1e^{-5} > x$	13,2%	13,0%	13,8%	13,3%	13,9%
Total gènes *Ath*	6 832	4 184	5 377	4 030	6 098

FIG. 3.1 – Répartition des *gènes similaires* sur les chromosomes d'*Arabidopsis*.

L'organisation des *gènes similaires* au sein du génome d'*Arabidopsis* a été étudiée. Les gènes ont été localisés à leur position respective sur chacun des chromosomes (Figure 3.1). Les 3 635 *gènes similaires* d'*Arabidopsis* ne semblent pas être regroupés ni présentés une

organisation particulière au sein du génome de cette plante. On retrouve environ 4% de gènes à *fortes similitudes* et 10% de gènes à *faibles similitudes* par chromosome. Ils sont répartis aléatoirement tout au long des 5 chromosomes en proportion similaire (à l'exception des régions centromériques pauvres en gènes et en gènes similaires). Toutefois, le chromosome 5 présente un pourcentage de gènes similaires un peu plus important que les autres, nous avons donc d'abord porté notre attention sur ce chromosome.

3.1.2 Le Génome d'*Arabidopsis* est-il Homéologues avec le Génome du Tournesol

Deux stratégies en parallèle

L'organisation du génome du tournesol et la synténie avec le génome d'*Arabidopsis* vont être analysées en utilisant deux stratégies complémentaires. L'une va se focaliser sur l'étude de la macro-organisation du génome en utilisant des marqueurs moléculaires et de la cartographie génétique et l'autre va se focaliser sur la micro-organisation du génome du tournesol en analysant les gènes du tournesol (éloignement, duplication, etc).

La première analyse consiste à étudier l'organisation de séquences EST ou ARNm au sein du génome du tournesol ainsi que de leur similitudes avec des gènes d'*Arabidopsis* (étude de la macro-organisation). Cette analyse va nous permettre d'avoir une vision globale de leur arrangement au niveau chromosomique et de le comparer à celui d'*Arabidopsis*.

La deuxième analyse consiste à étudier l'organisation des séquences d'EST ou ARNm de tournesol similaires à des gènes d'*Arabidopsis* distants en moyenne de 45 Kbases les uns des autres dans deux régions du chromosome 5 chez *Arabidopsis* (étude de la micro-organisation du génome). Cette éloignement entre les gènes d'*Arabidopsis* devrait permettre d'obtenir deux gènes par clones BAC de tournesol (80 kbases de moyenne) si l'éloignement entre gènes de tournesol est identique à l'éloignement observé chez *Arabidopsis*. Cette analye va nous permettre d'avoir une vision de l'organisation des gènes du tournesol au niveau d'une région particulière du génome et de la conservation de la micro-synténie avec *Arabidopsis*.

Quel que soit la méthode utilisée, les amorces et les sondes Overgo sont définies à l'aide d'lccare à partir de séquences EST ou ARNm de tournesol ayant de faibles ou de fortes similitudes avec des gènes d'*Arabidopsis*. Les séquences de tournesol sont sélectionnées en fonction des similarités avec des gènes d'*Arabidopsis*.

Pour l'étude de la macro-organisation, on utilise les séquences EST de tournesol qui ont servies à la construction de la carte génétique de Gentzbittel *et al.* de 1999 ainsi que les séquences EST ou ARNm de *fortes similitudes* qui sont régulièrement espacées tout au long du chromosome 5 chez la plante modèle (distance en moyenne de 546 Kbases). Des amorces spécifiques sont alors définies et celles qui présentent du polymorphisme de taille servent de marqueurs moléculaires qui seront intégrés dans la carte génétique de Rachid Al-chaarani *et al.* (2004).

Pour l'étude de la micro-organisation, on utilise des séquences EST ou ARNm de *faibles* et *fortes similitudes* qui sont toutes localisées à des positions les plus proches possibles les unes des autres sur le chromosome 5 de la plante modèle (distance en moyenne de 45 Kbases). Des sondes Overgo spécifiques de ces séquences EST ou ARNm de tournesol sont alors définies et utilisées pour cribler une banque de clones BAC (Gentzbittel *et al.*, 2002) afin d'identifier les clones BAC qui contiennent ces séquences. Les clones BAC positifs sont alors analysés pour tenter de les ordonner. Si la synténie est conservée entre le tournesol et *Arabidopsis* et que les sondes utilisées sont aussi proches les unes des autres, alors les clones BAC devraient pouvoir être ordonnés et la carte physique du génome du tournesol sera établie. Parmi les sondes Overgo, quelques unes sont définies à partir de séquences EST d'autres organismes que le tournesol pour combler les régions du chromosome 5 d'*Arabidopsis* qui ne présentent aucun gène similaire à des séquences EST ou ARNm. Ces sondes Overgo sont définies dans des régions conservées avec au moins 3 ou 4 organismes (dont *Arabidopsis*).

Ces différentes manipulations vont nous permettre d'évaluer l'efficacité d'amplification PCR sur le tournesol à partir des couples d'amorces définies de part et d'autre des introns et de la méthode de criblage de la banque de clones BAC (sondes issues de séquences de tournesol ou d'autres organismes). L'analyse de la synténie entre le génome du tournesol et de celui d'*Arabidopsis* dépend des résultats obtenus pour la construction de la carte physique et de la carte génétique du tournesol. Si la synténie est conservée entre tournesol et *Arabidopsis* alors les sondes Overgo devraient permettre d'ordonner les clones BAC entre eux, et les amorces devraient permettre, sous réserve de polymorphisme, de construire une carte génétique qui présentera des marqueurs liés. Les résultats obtenus devraient, de ce fait, confirmer ou infirmer la conservation de synténie entre les deux organismes.

3.2 Résultats expérimentaux

3.2.1 Marqueurs Moléculaires et Cartes Génétiques

Dans un premier temps nous avons comparé les séquences EST et ARNm de tournesol qui ont servi à la construction de la carte génétique de Gentzbittel *et al.* de 1999. Un total de 121 séquences EST utilisées pour la cartographie génétique a été comparé aux séquences codantes des gènes d'*Arabidopsis thaliana* par Iccare. Seules 41 séquences présentent des similitudes avec 38 gènes d'*Arabidopsis* (les résultats sont accessibles à cette adresse http://bioinfo.genopole-toulouse.prd.fr/Iccare/pers_plante/vse_20Jun2005_09_40_55/).
Le pourcentage de séquences EST similaires aux gènes d'*Arabidopsis* est de 34% ce qui correspond au même pourcentage de similitude que l'ensemble des séquences EST et ARNm similaires aux gènes d'*Arabidopsis*.

Certaines des séquences EST de tournesol de la carte génétique de Gentzbittel *et al.* (1999) ont été cartographiées à plusieurs loci. Ces séquences EST doivent correspondre à des gènes dupliqués ou appartenant à des familles multigéniques. Nous avons donc regardé parmi les

gènes similaires d'*Arabidopsis* ceux qui étaient dupliqués ou appartiennent à des familles multigéniques avec Synteny Search. Parmi les 38 *gènes similaires* d'*Arabidposis*, 10 sont uniques ou présentent de très faible similitudes avec d'autres gènes d'*Arabidopsis* (supérieur à $1e^{-15}$) et 28 sont dupliqués ou appartiennent à une famille multigénique. Sur les 10 gènes uniques, 7 sont similaires à une séquence EST localisée à un seul locus et 3 sont similaires à une séquence EST localisée à deux loci. Sur les 28 gènes dupliqués ou appartenant à une famille mutligénique, 10 gènes sont similaires à une séquence EST localisée à plusieurs loci et 16 gènes sont similaires à une séquence localisée à un seul locus (2 des gènes n'ont pu être déterminés).

Ces données laissent à penser que le nombre de gènes présent chez le tournesol semble proche du nombre de gène présent chez *Arabidopsis*. Les gènes uniques chez *Arabidopsis* sont majoritairement retrouvés unique (en cartographie génétique) chez le tournesol et ceux dupliqués ou appartenant à une famille multigénique sont retrouvés pour plus d'un tiers cartographiés à plusieurs loci.

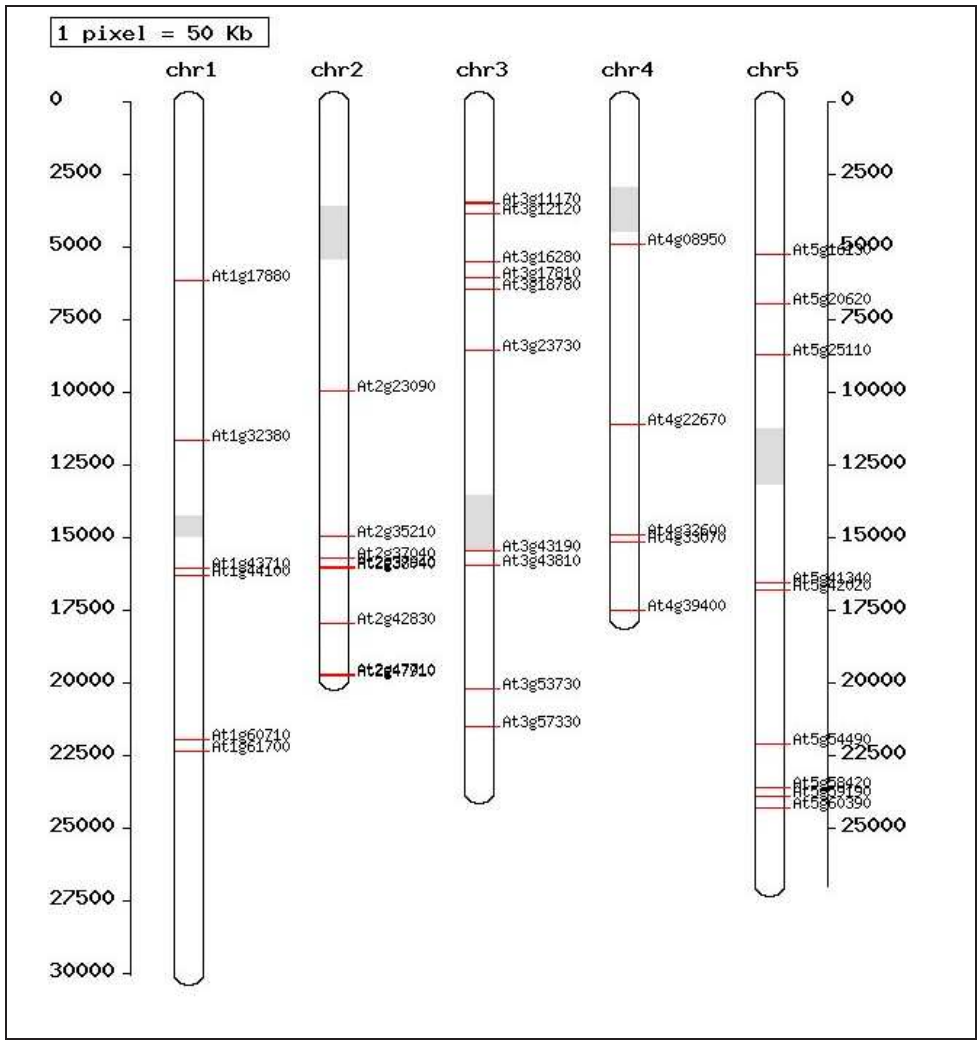


FIG. 3.2 – Position des gènes d'*Arabidopsis* similaires aux séquences de tournesol cartographiées.

La position des *gènes similaires* d'*Arabidopsis* est présentée dans la Figure 3.2, et la position des séquences EST sur la carte génétique du tournesol est présentée dans l'annexe *G* page 170. La comparaison des données cartographiques entre les gènes similaires d'*Arabidopsis* et les séquences du tournesol ne permet pas d'établir de régions conservées entre les deux espèces. Le nombre de gènes et de séquences EST similaires est trop faible et l'écart entre ceux-ci est trop grand pour établir de la macrosynténie entre le génome d'*Arabidopsis* et du tournesol.

Les séquences EST de la carte génétique de Gentzbittel *et al.* (1999) n'ayant pas permis de mettre en évidence la macrosynténie entre le génome d'*Arabidopsis* et celui du tournesol, nous avons décidé d'utiliser les séquences EST ou ARNm du tournesol fortement similaires à des gènes d'*Arabidopsis* localisés sur le chromosome 5 et distant en moyenne de 546 Kbases pour créer de nouveaux marqueurs moléculaires. Des couples d'amorces sont définis de part et d'autre d'un intron pour chacune des séquences EST qui présentent de fortes similitudes avec les gènes de la plante modèle. Les produits d'amplicification qui présentent du polymorphisme sont alors utilisés comme marqueurs moléculaires et sont intégrés à la carte génétique de Rachid Al-Chaarani *et al.* de 2004 (ces marqueurs ne pouvant malheureusement pas être intégreés à la carte de Gentzbittel *et al.* de 1999).

Les couples d'amorces définies à partir de séquences EST ou ARNm de tournesol de *fortes similitudes* ont été testés sur différents cultivars de tournesol. Les amorces 1, 2, et 4 à 11 (soit 10 couples d'amorces) ont été testées sur 2 ou 3 ADN différents (PAC2, RHA266 et psc8) et les résultats de migration des produits PCR sont présentés dans la Figure 3.3. Les deux ou trois cultivars utilisés sont amplifiés pour les 10 couples d'amorces utilisés, à l'exception des cultivars PAC2 et psc8 qui n'ont pas été amplifiés respectivement par le couple d'amorces 11 et le couple d'amorces 4. Les amorces 1, 2, 3, et 11 à 51 (soit 44 couples d'amorces) ont été testées sur 5 cultivars de tournesol (ordre de dépôt dans les puits de gauche à droite pour chaque couple d'amorces : AS613, SD1', psc8, PAC2, RHA266) et les résultats de la migration sont présentés dans la Figure 3.4. Parmi ces 44 couples d'amorces, 5 n'amplifient aucun cultivar (39, 48, 49, 50 et 51) et 39 couples d'amorces amplifient au moins 1 cultivar. Parmi les amorces qui fonctionnent, 3 d'entre elles sont très faiblement amplifiées (16, 28 et 36). Au final, 43 couples d'amorces amplifient tous les cultivars, 3 sont faiblement amplifiés et 5 ne fonctionnent pas.

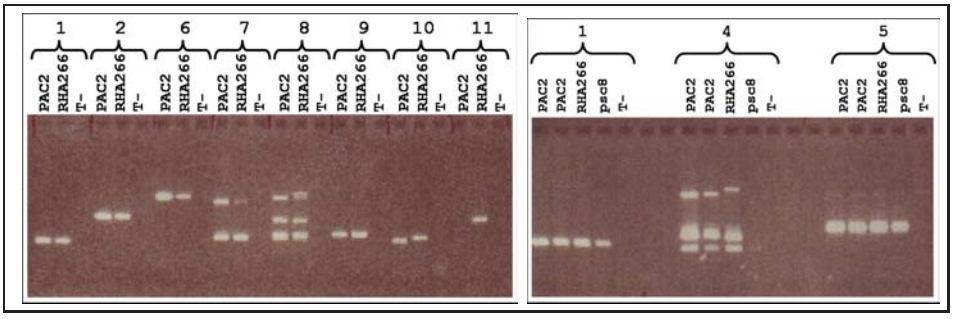


FIG. 3.3 – Migration des produits d'amplification des amorces définies à partir de séquences EST de tournesol similaires sur 2 ou 3 cultivars de tournesol.

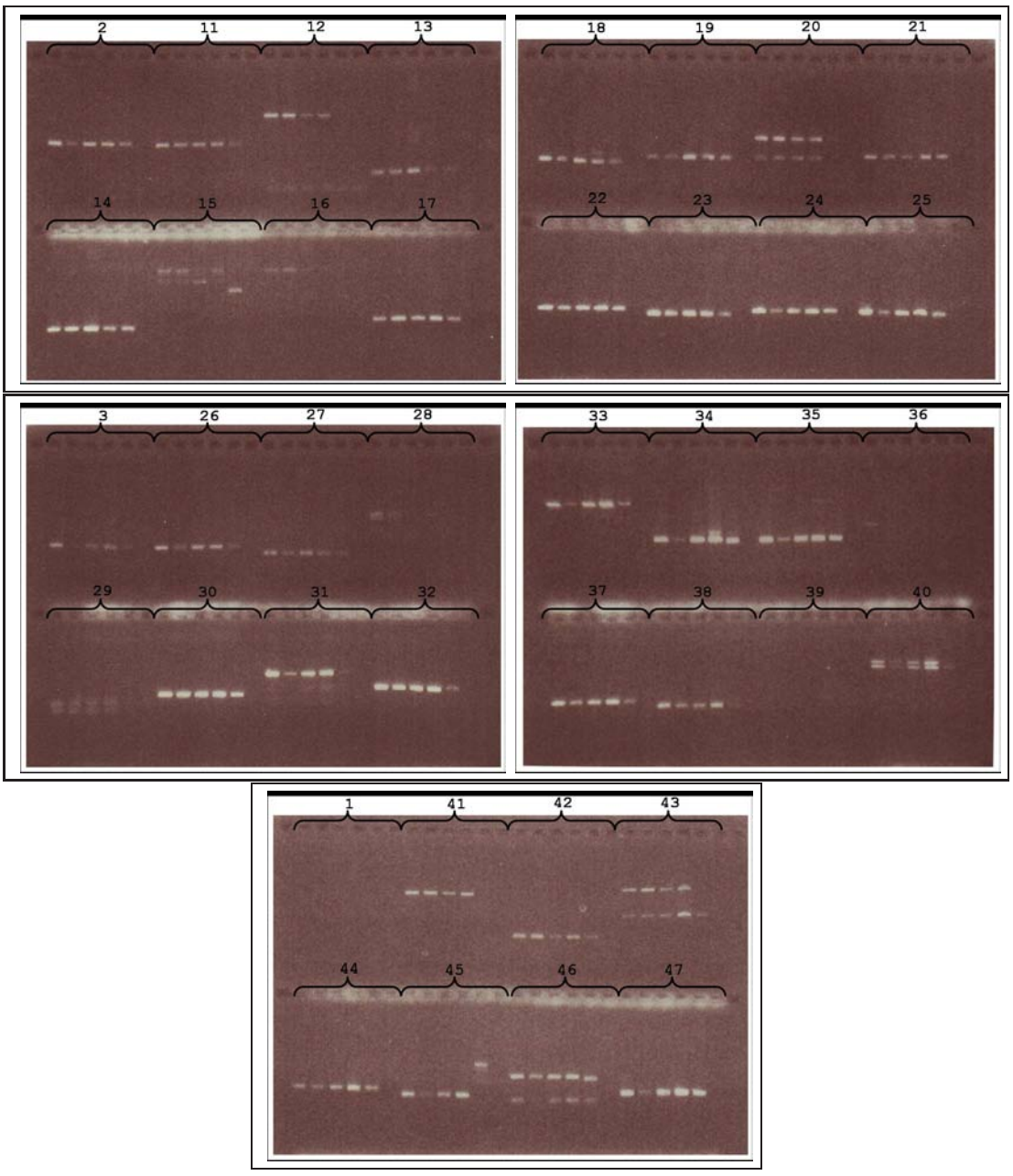


FIG. 3.4 – Migration des produits d'amplification des amorces définies à partir de séquences EST de tournesol similaires sur 5 cultivars de tournesol.

Certains couples d'amorces amplifient des fragments qui présentent un polymorphisme de taille en fonction du cultivar. Ce polymorphisme va être exploité afin de faire des marqueurs moléculaires. Ainsi, parmi les 46 couples d'amorces qui fonctionnent, 4 présentent un polymorphisme de taille entre le cultivar PAC2 et RHA266 (4, 10, 15 et 45) sur gel d'agarose 2%, 4 autres semblent présenter du polymorphisme de taille entre PAC2 et RHA266 (car difficilement discernable sur gel d'agarose 2% : 18, 25, 32 et 46) et 11 couples d'amorces présentent un polymorphisme de type présence/absence (8, 11, 12, 20, 29, 31, 34, 38, 40, 41 et 43). Les ADN des lignées recombinantes ont été amplifiés, ainsi que les ADN de PAC2 et RHA266 (parents des lignées recombinantes) par 8 couples d'amorces (4, 8, 10, 15, 25, 40, 43 et 45). Les produits d'amplification ont été séparés sur gel d'agarose à 4% et les résultats de ces migrations sont présentés dans l'annexe *H* page 172. Les produits d'amplification du couple d'amorces 43 ne présentent pas de polymorphisme de taille ni de présence/absence entre PAC2 et RHA266 (résultats non présentés). Pour les autres couples d'amorces testés, il existe bien un polymorphisme de taille ou de présence/absence. La migration des produits d'amplification issus des couples d'amorces 4 et 15 sont plus difficiles à analyser comparés aux autres couples d'amorces. Sur les 8 marqueurs moléculaires potentiels testés, 7 présentent effectivement un polymorphisme de taille ou d'absence/présence et un marqueur n'en présente pas. Le polymorphisme repéré sur gel d'agarose 2% est vérifié pour 90% des couples d'amorces testés.

Sur les 8 couples d'amorces testés, sept (4, 8, 10, 15, 25, 40 et 45) présentent du polymorphisme entre PAC2 et RHA266, ils sont utilisés comme marqueurs moléculaires et passés sur 93 individus de la population de RIL. Ces marqueurs moléculaires sont notés respectivement P4, P8, P10, P15, P25, P40 et P45. Les différents marqueurs sont notés en fonction du profil parental : profil PAC2 noté *A* et profil RHA266 noté *B* pour chacune des lignées recombinantes. Les données des marqueurs moléculaires sont regroupés dans un fichier et sont aussi combinées aux informations de cartographie qui ont servi à la construction de la carte génétique de Rachid Al-Chaarani *et al.* (2004). Ces fichiers sont ensuite traités par MapMaker.

Pour le fichier qui ne contient que nos marqueurs moléculaires (P4, P8, P10, P15, P25, P40 et P45), MapMaker ne trouve aucune liaison entre les marqueurs moléculaires avec une valeur de LOD de 3.0 ou de 2.0. Les marqueurs ne sont donc pas liés entre eux. Nos marqueurs sont ensuite combinés à ceux utilisés pour la construction de la carte génétique de Rachid Al-Chaarani *et al.* (2004) qui est constituée de 411 marqueurs (367 marqueurs originels et 44 nouveaux marqueurs ajoutés depuis la publication) formant 21 groupes de liaisons. Les liaisons entre marqueurs sont déterminées par MapMaker avec une valeur de LOD équivalente à celle qui a servi à la construction de la carte publiée (LOD de 4.0, résultats présentés dans l'annexe *I* page 175). Un total de 33 groupes de liaisons est établi. Huit d'entre eux sont de nouveaux groupes de liaisons composés de nouveaux marqueurs (entre 2 et 5 marqueurs). Les 25 autres groupes de liaisons correspondent aux 21 groupes établis dans la publication. Dix-sept de ces groupes correspondent aux groupes définis dans la carte publiée avec 1 ou 2 nouveaux marqueurs, les 8 derniers groupes correspondent aux 4 groupes restants. Ainsi les groupe de liaisons 7 et 23 correspondent au groupe de liaisons 6 de la carte publiée. La correspondance entre groupes de liaisons définis par MapMaker pour notre analyse et les groupes

de liaisons de la carte publiée est présentée dans le Tableau 3.1.

Groupe MapMaker	groupe carte publiée	Groupe MapMaker	groupe carte publiée
1	1	12	28
2	2	13	9
3	13	14	17+26
4	5	15	16+22
5	6	16	8
6	7+23	17	20+18
7	14	18	24
8	19	19	10
9	11	20	15
10	3	21	21
11	12		

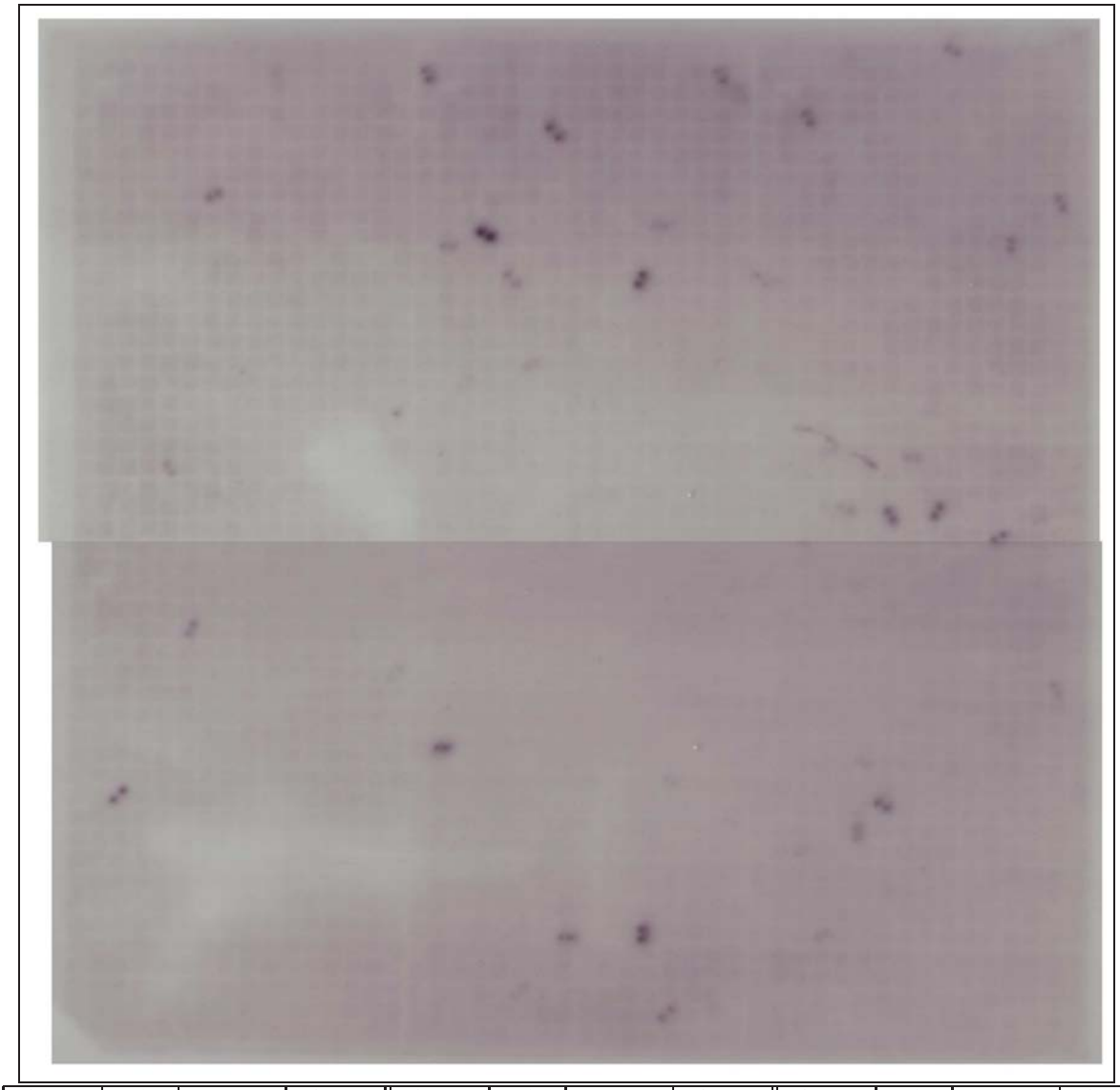
TAB. 3.1 – Correspondances entre les groupes de liaisons de notre analyse et ceux de la carte publiée.

Sur nos 7 marqueurs, P8, P15 et P40 (respectivement marqueurs 412, 414 et 416 dans l'analyse de MapMaker) ne sont pas liés à la carte. Le marqueur P04 (418) se retrouve à l'extrémité du groupe de liaisons 2. Le marqueur P10 (413) se retrouve entre les marqueurs 339 et 171 dans le groupe de liaisons 10. Le marqueur P25 (415) est fortement lié aux marqueurs 365 et 324 dans le groupe de liaisons 15. Le marqueur P45 (417) se retrouve à l'extrémité du groupe de liaisons 17 (pour plus de détails, voir l'annexe *I* page 175).

L'espacement entre les gènes d'*Arabidopsis* du chromosome 5 qui présentent des similitudes avec des séquences EST de tournesol est d'environ 546 Kbases sur la carte physique et d'environ 2 cM sur la carte génétique. Sur les 51 couples d'amorces définis, 46 couples d'amorces ont amplifiés des cultivars, 8 couples d'amorces ont été testés pour le polymorphisme et seul 4 marqueurs moléculaires ont pu être cartographiés chez le tournesol. L'espacement entre ces gènes similaires d'*Arabidopsis* est respectivement de 3 Mbases entre P04 et P10, de 9 Mbases entre P10 et P25 et de 10 Mbases entre P25 et P45. Chez le tournesol, ces 4 marqueurs moléculaires qui ont pu être cartographiés (P04, P10, P25 et P45) sont sur des groupes de liaisons différents donc non liés. Ces résultats ne permettent pas d'établir de la macrosynténie entre le génome d'*Arabidopsis* et celui du tournesol. Si la conservation de synténie entre le génome du tournesol et celui d'*Arabidopsis* existe, le génome des 2 plantes a du subir de nombreux remaniement de blocs chromosomiques d'une taille inférieure à 3 Mbases (au niveau du génome d'*Arabidopsis*).

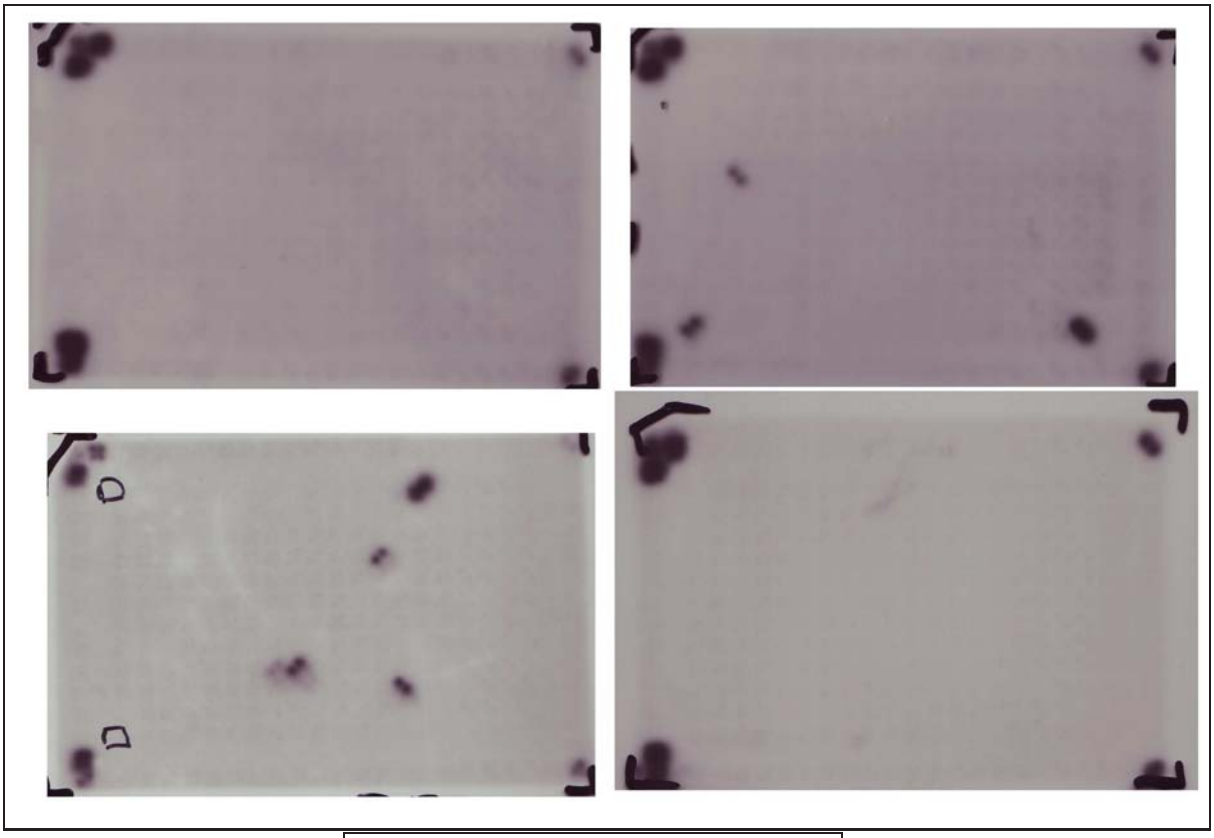
3.2.2 Criblage de la Banque de Clones BAC

La banque de clones BAC a été criblée en utilisant des pools de 48, 96 ou 149 sondes et chaque sonde a été hybridée au moins deux fois sur deux lots différents de membranes. Au total, 161 sondes Overgo ont servi à cribler la banque de clones BAC. La Figure 3.5 présente le résultat de l'hybridation de 149 sondes (sondes n° 1 à 151) sur la membrane *F* après 5 jours et demi d'exposition à -80°C. Les 6 champs apparaissent clairement sur la membrane. L'analyse des résultats se fait à l'aide du pattern de dépôt (voir l'annexe *E* page 164). Chaque clone BAC positif peut être associé à un clone localisé sur une plaque et dans un puits. Le



champs	puits	patterns	plaques	champs	puits	patterns	plaques	champs	puits	patterns	plaques
1	C12	6	246	4	FO9	4	268	5	C10	4	276
1	P18	6	246	4	D15	3	267	5	F11	2	274
1	F20	8	248	4	K13	8	272	5	O18	7	279
2	G17	4	252	4	C08	6	270	6	I02	8	288
2	E04	7	255	4	E13	5	269	6	L01	4	284
3	B15	1	257	4	G20	5	269	6	M15	2	282
3	F03	6	262	4	L16	1	265	6	O17	7	287
3	H06	1	257	4	O23	7	271	6	J24	3	283
3	M02	6	262	4	B15	1	265	6	E02	3	283
3	M13	3	259	5	D06	4	276	6	H05	1	281
3	L06	2	258	5	F14	3	275	6	C21	7	287
3	P11	3	259	5	G12	3	275	6	G02	7	287
4	A23	7	271								

FIG. 3.5 – Résultats de l'hybridation de 149 sondes Overgo sur la membrane F après 5 jours et demi d'exposition à -80°C.



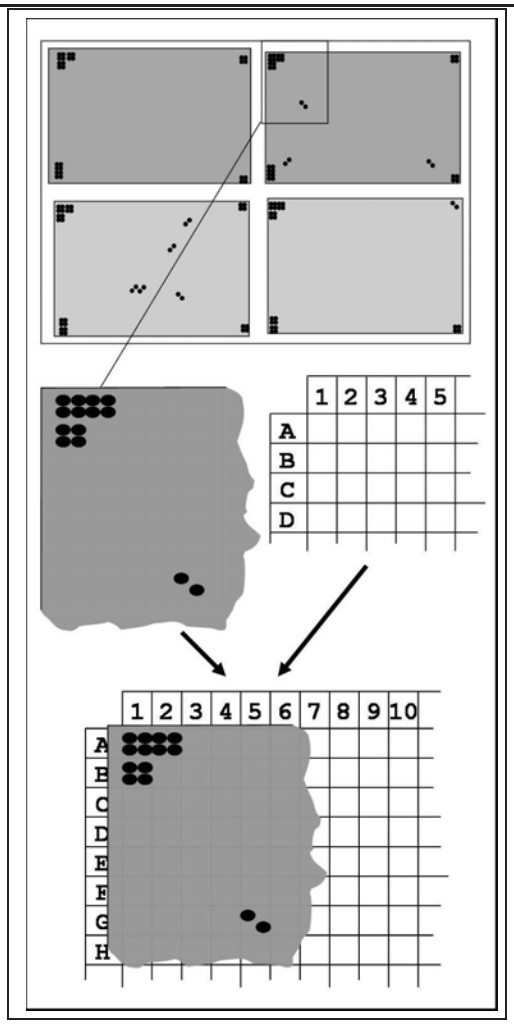


FIG. 3.6 – Résultats de l'hybridation de 4 sondes Overgo sur les nouvelles membranes.

tableau de la Figure 3.5 présente les résultats de l'analyse d'un pool de 149 sondes (n°1 à 151) sur cette membrane F. Les autres membranes sont analysées de la même façon et l'ensemble des sondes sont utilisées pour cribler la banque de clones BAC. Les 161 sondes Overgo ont permis d'isoler 503 clones BAC.

Les clones BAC positifs ont été repiqués (à l'exception de 12 clones BAC qui n'ont pas poussé) et de nouvelles membranes ont été faites à partir de ces clones BAC. Les sondes ont été hybridées individuellement sur les nouvelles membranes. La Figure 3.6 présente les résultats des hybridations pour 4 sondes, une sonde par membrane (en plus de la sonde Rub2). En haut à gauche se trouve la membrane correspondant à la sonde r8 spécifique de la rubisco, puis en suivant le sens des aiguilles d'une montre on retrouve les résultats de la sonde 81, ensuite ceux de la sonde 145 et enfin ceux de la sonde 122. Ayant déjà été analysés, les clones BAC spécifiques de la rubisco avaient déjà été récupérés, ceux-ci ont été utilisés pour repérer le sens et l'orientation des membranes. Les sondes de rubisco sont utilisées sur chaque membrane en plus des sondes individuelles afin de faciliter la lecture des membranes. Les 3 clones BAC en haut à gauche permettent de positionner les puits A01, B01 et A02, les clones positifs en bas à droite permet de positionner le puits P24. La sonde rubisco n'est positive qu'avec les clones BAC des 4 coins de la membrane (puits A01, A02, B01, O01, P01, A24 et P24). Ceci permet de faciliter l'analyse des membranes en positionnant les puits A01 et P24 et éviter ainsi les erreurs d'interprétation lors de l'analyse.

La sonde 81 s'hybride avec les clones BAC 234K15, 29D15 et 208G04 (respectivement puits N03 pattern 2, G05 pattern 1 et N21 pattern 1). La sonde 145 ne s'hybride avec aucun clone BAC. La sonde 122 s'hybride avec 5 clones BAC, 312C01, 325C23, 377F13, 153N21 et 383A07 (respectivement K10 pattern 2, K11 pattern 2, F15 pattern 2, L16 pattern 1 et C17 pattern 2).

Les 161 sondes sont donc passées individuellement sur les membranes. Ces hybridations sonde à sonde ont permis d'identifier tous les clones BAC spécifiques d'une sonde (résultats présentés dans l'annexe F page 167). Au total, 330 clones BAC ont pu être associés aux 161 sondes Overgo utilisées.

Le dépoolage des sondes a permis d'identifier 330 clones BAC sur les 491 récupérés (12 n'ont pas poussé) ce qui fait un taux de vérification de 68% entre la première hybridation (récupération des clones BAC en pool) et la seconde (identification des clones BAC par sonde). La seconde hybridation identifie un nombre un peu plus faible de clones BAC que lors de la récupération. Nous avons alors regardé si la différence entre clones récupérés et clones identifiés est identique pour les 9 membranes de la banque de clones BAC pour les sondes 1 à 151, soit 149 sondes. Les résultats sont présentés dans le Tableau 3.2. Les membranes E, G, H et I présentent le plus fort taux d'identification (supérieur à 75%), les membranes A, C, D et F présentent un taux un peu plus faible compris entre 63 et 70%. La membrane B présente un taux très faible de 43%. La qualité des membranes (quantité d'ADN, traitement des membranes) a une influence sur l'efficacité d'identification des clones BAC.

mbre	Hyb 1	Hyb 2	%
A	47	30	64%
B	58	25	43%
C	54	37	68%
D	49	34	69%
E	45	35	78%
F	48	30	63%
G	54	46	85%
H	51	42	82%
I	20	15	75%

TAB. 3.2 – Nombre de clones BAC positifs puis identifiés lors de l'hybridation de 149 sondes Overgo en fonction des membranes.

Nous nous sommes aussi intéressés à l'efficacité des sondes Overgo en fonction des similitudes existant entre la séquence EST ou ARNm de tournesol utilisée comme matrice pour définir la sonde Overgo et les gènes d'*Arabidopsis*. L'efficacité d'hybridation des sondes Overgo est constatée en fonction de la similitude entre séquences (e^{-20}, e^{-05} ou *multi*). Les résultats sont présentés dans le Tableau 3.3. Les sondes Overgo e^{-20} définies à partir de séquences de *fortes similitudes* s'hybrident dans 3 cas sur 4. Les sondes Overgo e^{-5} qui sont définies à partir de séquences de *faibles similitudes* présentent un taux d'efficacité un peu plus faible, un peu moins de 2 sur 3. Les sondes Overgo *multi* définies à partir de séquences d'autres organismes que le tournesol présentent un très faible taux d'efficacité, un peu plus de 1 sur 4.

Les sondes Overgo présentent un fort taux d'efficacité (supérieur à 60%) lorsqu'elles sont définies à partir de régions conservées des séquences EST ou ARNm de tournesol. L'efficacité des sondes Overgo est aussi influencée par la similitude entre les séquences de tournesol et d'*Arabidopsis* ; plus la similitude entre la séquence de tournesol et celle d'*Arabidopsis* est forte plus la sonde Overgo définie à partir de la séquence de tournesol est efficace. Les sondes définies à partir de séquences autres que tournesol présentent une efficacité beaucoup plus faible malgré la conservation de similitudes entre les séquences utilisées.

Type de sondes	e^{-20}	e^{-5}	*multi*
Sondes qui s'hybrident	83	21	3
Sondes qui ne s'hybrident pas	32	14	8
% d'efficacité	72%	60%	27%

TAB. 3.3 – Efficacité des sondes Overgo en fonction du type de sondes.

Par la suite nous nous sommes intéressés au nombre de clones BAC spécifiques de chaque sonde. Les sondes e^{-20} et e^{-5} ont été classées en fonction du nombre de clones BAC spécifiques de chaque sondes (les sondes *multi* n'ont pas été utilisées car seuls 27% s'hybrident avec des clones BAC). La moyenne du nombre de clones BAC par sonde est de 3,56 (sans tenir compte des sondes qui ne se sont pas hybridées). La distribution du nombre de clones BAC positifs par sonde a été analysée. Dix catégories correspondant au nombre de clones BAC positifs avec une sonde ont été établies. Les résultats obtenus pour les sondes e^{-20} et les sondes e^{-5}

sont identiques (résultats non présentés), elles ont donc été analysées ensemble. La Figure 3.7 présente les résultats de cette analyse. Environ 30% de sondes ne s'hybrident avec aucun clone BAC (cummul des résultats entre l'efficacité des sondes Overgo e^{-20} à 72% et celles e^{-5} à 60%) et 38% des sondes s'hybrident avec 1 ou 2 clones BAC. Lorsque l'on regarde uniquement les sondes qui s'hybrident alors le pourcentage de sondes qui s'hybrident avec 1 ou 2 clones BAC passe à 56% et il est de 72% en incluant les sondes qui s'hybrident avec 3 clones BAC. Il semble donc que les sondes s'hybrident majoritairement avec 2 ou 3 clones BAC et que la moyenne d'hybridation est un peu trop élevée.

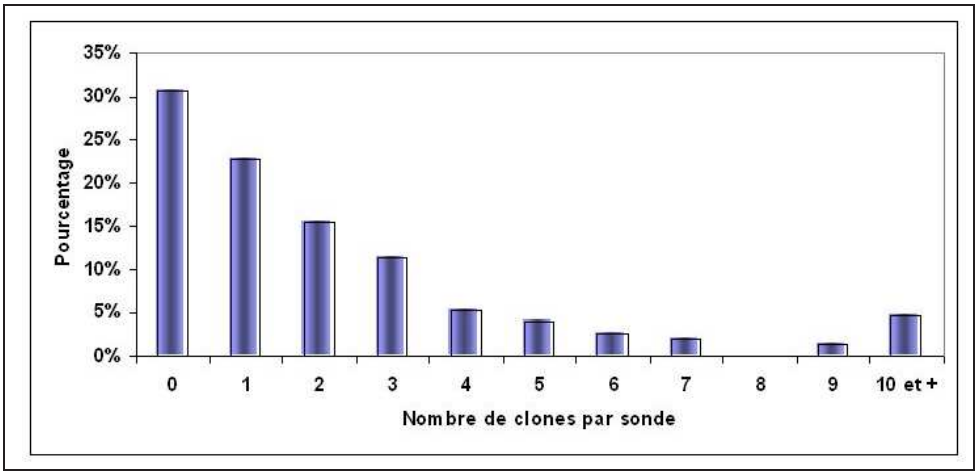


FIG. 3.7 – Distribution du nombre de clones BAC positifs par sonde Overgo.

La différence de résultats d'hybridation entre les sondes qui ne s'hybrident qu'avec 2 à 3 clones BAC et celles qui s'hybrident avec plus de 10 clones BAC s'explique par la non spécificité stricte des sondes Overgo pour un seul gène. Les sondes Overgo ont été définies dans des régions conservées des séquences de tournesol. Les gènes dupliqués ou appartenant à des familles multigéniques peuvent toutt aussi bien avoir cette région conservée pour plusieurs gènes. De ce fait les sondes Overgo utilisées ne seront pas spécifique d'un gène mais d'une famille de gène. Pour tenir compte du fait que certaines sondes Overgo sont spécifiques de plusieurs gènes, nous avons analysé les gènes d'*Arabidopsis*, qui présentent des similitudes avec les séquences EST ou ARNm utilisées pour définir des sondes Overgo, à l'aide de Synteny Search pour déterminer si ces gènes sont uniques, dupliqués ou présentent des similitudes avec d'autres gènes d'*Arabidopsis*.

Les gènes d'*Arabidopsis* présentant des similitudes avec les séquences EST ou ARNm de tournesol sont tous localisés que le chromosome 5. Le chromosome 5 d'*Arabidopsis* présente de nombreuses régions qui sont dupliquées. La région du chromosome 5 d'*Arabidopsis* situé de 0 à 7 000 kbases (haut du chromosome 5) est consitué de plusiers régions chromosomiques qui sont le résultat d'évènement de duplications du chromosome 3 mais aussi du chromosome 5 lui-même (Figure 3.8). Ainsi, les gènes situés sur le chromosome 5 dans les régions 0 - 2 150 kbases, 4 500 - 6 150 Kbases et 6 150 - 6 500 Kbases sont respectivement dupliqués avec les gènes localisés sur le chromosome 3 dans les régions 2 650 - 3 950 kbases, 0 - 1 200 Kbases et 1 800 - 1 950 kbases (640, 474 et 79 gènes localisés dans chaque région, 113, 116 et 21 sont

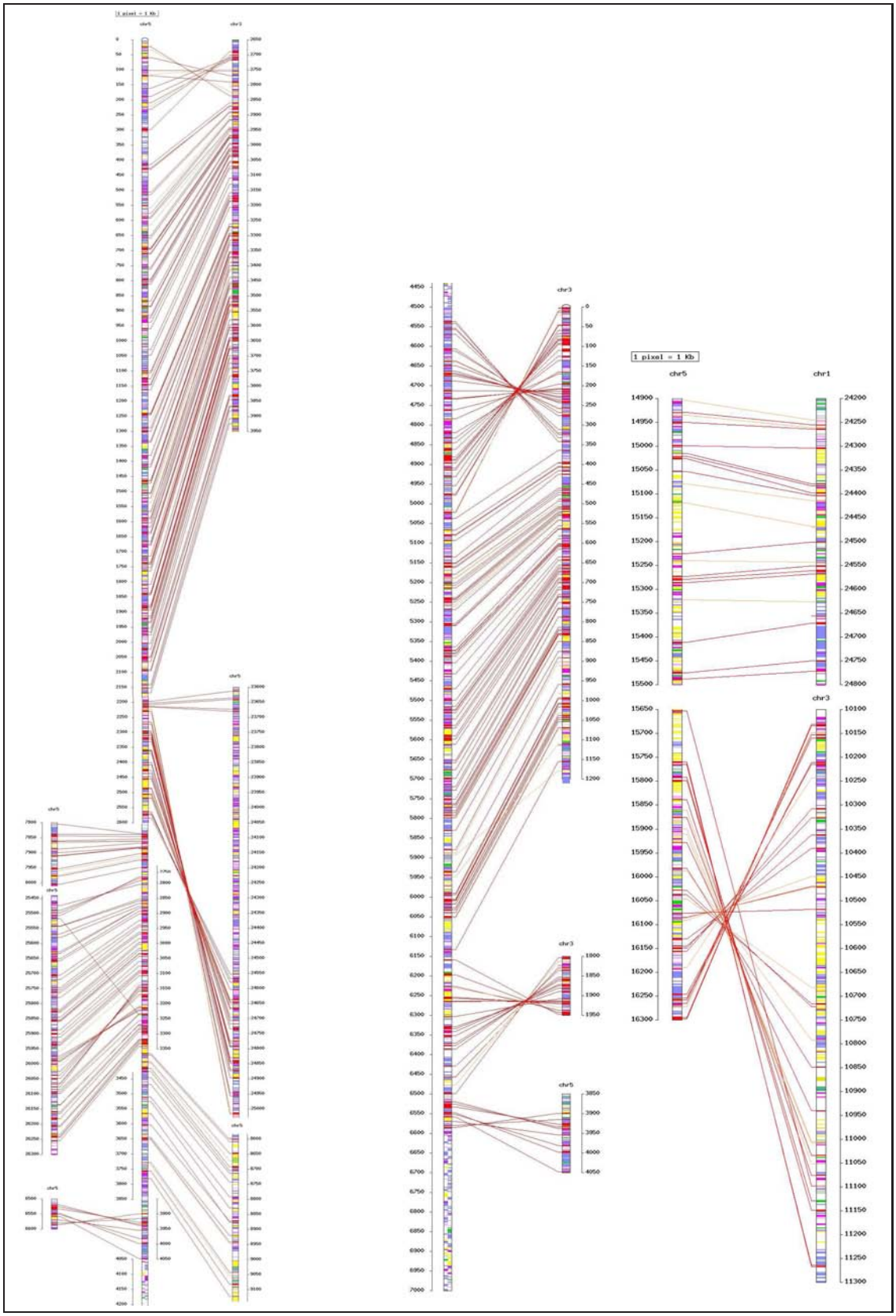


FIG. 3.8 – Duplication des gènes d'*Arabidopsis* pour la région 0-7 Mbases et 14,9-16,3 Mbases du chromosome 5 avec le reste du génome.

dupliqués pour chaque région, 285, 198 et 30 sont uniques pour chaque région et 242, 160 et 28 sont dupliqués mais avec d'autres régions du génome pour chaque région) ; de même avec les gènes du chromosome 5 des régions 2 150 - 2 600 kbases, 2 600 - 2 750 kbases, 2 750 - 3 350 kbases, 3 350 - 3 850 kbases et 3 850 - 4 050 kbases qui sont dupliqués avec les gènes des régions 23 600 - 25 050 kbases, 7 800 - 8 050 kbases, 25 450 - 26 300 kbases, 8 600 - 9 200 kbases et 6 500 - 6 600 kbases de ce même chromosome 5 (118, 42, 170, 135 et 53 gènes localisés dans chaque région, 39, 15, 54, 19 et 10 sont dupliqués pour chaque région, 36, 20, 56, 65 et 26 sont uniques pour chaque région et 42, 7, 60, 51 et 17 sont dupliqués mais avec d'autres régions du génome pour chaque région). La région du chromosome 5 contenant les gènes de la rubisco (14 900 - 16 300 kbases) présente aussi des duplications, ainsi les gènes situés sur le chromosome 5 dans la région 14 900 - 15 500 kbases sont respectivement dupliqués avec les gènes localisés sur le chromosome 1 dans la région 24 200 - 24 800 kbases (126 gènes localisés dans cette région, 20 sont dupliqués, 34 sont uniques et 84 sont dupliqués mais avec d'autres régions du génome) ; de même avec les gènes du chromosome 5 de la région 15 650 - 16 300 kbases qui sont dupliqués avec les gènes de la région 10 100 - 11 300 kbases du chromosome 3 (169 gènes localisés dans cette région, 32 sont dupliqués, 51 sont uniques et 94 sont dupliqués mais avec d'autres régions du génome).

Au total, 200 gènes d'*Arabidopsis* similaires à des séquences de tournesol sont passés au crible avec Synteny Search (12 gènes dans la région de la rubisco, 149 dans la région 0 - 7 Mbases du chromosome 5 et 39 pour les couples d'amorces). Parmi ces 200 gènes d'*Arabidopsis*, 67 de ces gènes sont uniques et 133 sont dupliqués ou appartiennent à des familles multigéniques chez *Arabidopsis*.

Pour tenir compte du fait que chez le tournesol, il doit aussi y avoir des gènes dupliqués ou appartenant à des familles multigéniques, nous avons considéré que des gènes uniques chez *Arabidopsis* l'étaient aussi chez le tournesol, de même pour les gènes dupliqués ou appartenant à des familles multigéniques. A partir de là nous avons mis en parallèle le nombre de bandes visibles sur gel d'agarose ou le nombre de clones BAC positifs à une sonde avec le nombre de gènes d'*Arabidopsis* similaires aux séquences EST de tournesol. Ce travail a été réalisé pour 94 sondes Overgo et 10 couples d'amorces (voir annexe *J* page 186).

Chez *Arabidopsis*, parmi ces 94 gènes d'*Arabidopsis*, 32 sont uniques et 62 sont dupliqués ou appartiennent à des familles multigéniques. Le nombre de clones BAC positifs aux sondes définies à partir des séquences EST ou ARNm similaires à nos 94 gènes d'*Arabidopsis* est divisé par le nombre de gènes similaires chez *Arabidopsis* (divisé par 1 s'il s'agit d'un gène unique, par 2 s'il est dupliqué, par 3 s'il est similaire à deux autres gènes, etc.). La moyenne du ratio "nombre de clones BAC positifs" par "gène potentiel" pour chacune des données (et non plus par sonde comme dans le paragraphe page 99) est de 1,8. La répartition des ratio "nombre de clones BAC positifs" par "gène potentiel" est présentée dans la Figure 3.9. Environ 28% des 94 gènes ne sont associés à aucun clone BAC et 46% sont associés à 1 ou 2 clones BAC. Lorsque l'on regarde uniquement ceux qui s'hybrident, environ 63% des gènes sont associés à 1 ou 2 clones BAC et on monte jusqu'à 84% avec 3 clones BAC. Ces résultats laissent à penser que la banque de clones BAC n'est pas 4 à 5 fois couvrante mais plutôt 1 à 2 fois couvrantes pour la région étudiée.

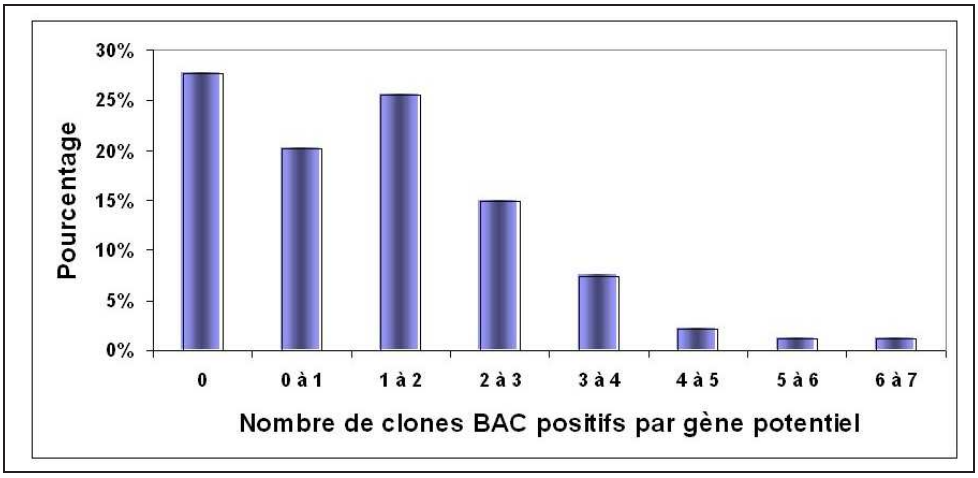


FIG. 3.9 – Distribution du nombre de clones BAC positifs par gène potentiel.

Après avoir regardé l'efficacité des sondes Overgo et la couverture de la banque de clones BAC, nous nous sommes intéressé aux clones BAC qui ont été hybridés par les sondes Overgo contigus chez *Arabidopsis*. A partir des sondes qui se sont hybridés avec des clones BAC, nous avons regardé si certains clones BAC étaient communs à plusieurs sondes afin de constuire une carte physique.

Pour les sondes de la région de la rubisco (r1 à r12) chez *Arabidopsis*, tous les clones BAC sont spécifiques d'une seule sonde pour le tournesol. Il est donc impossible des regroupés les clones BAC par sonde commune et donc de constuire une carte physique de la région.

En revanche, pour les sondes de l'autre région (sonde 1 à 151) du chromosome 5 chez *Arabidopsis*, lorsque l'on regarde les clones BAC associés à chaque sonde on s'aperçoit que certains clones BAC sont communs à plusieurs sondes. Ainsi les sondes 94 et 95 présentent 4 clones BAC communs, les sondes 7 et 136 présentent 3 clones BAC communs et les sondes 2 et 3 présentent un clone BAC commun. On peut donc regrouper les clones BAC des sondes 94 et 95 ensemble, ainsi que que ceux des sondes 7 et 136 et celui des sondes 2 et 3. Pourtant, la carte physique ne peut être construite avec ces clones BAC, car l'analyse avec Synteny Search révèle que les gènes utilisés pour définir les sondes 94 et 95 sont dupliqués, idem pour les sondes 7 et 136 et les sondes 2 et 3. Il est donc normal de récupérer les mêmes clones BAC avec ces sondes. D'autres clones BAC sont communs à plusieurs sondes, c'est le cas du clone $269E13$ commun aux sondes 17 et 101, du clone $289D05$ commun aux sondes 30 et 58, du clone $289L01$ commun aux sondes 39 et 150, du clone $399E03$ commun aux sondes 64 et 69, ou du clone $34F18$ commun aux sondes 51 et 52. Les différentes sondes utilisées ne présentent pas de relation entre elles lorsqu'on utilise Synteny Search, cependant à l'excpetion des sondes 51 et 52, toutes les autres ne sont pas proches chez *Arabidopsis*. Ces clones communs sont peut être des artéfacts d'hybridation ou des pseudogènes présent dans le génome du tournesol. Certains clones BAC sont parfois communs à plusieurs sondes qui elles-mêmes présentent en commun différents clones BAC formant des relations complexes entre sondes et clones BAC. Par exemple, les sondes 66 et 9 présentent quatre clones BAC communs ($151N12$, $159E22$, $171B09$ et $183L02$). Le clone $151N12$ est aussi commun à la sonde 84, alors que les clones

159E22 et 171B09 sont communs aux sondes 68 et 36 et aussi respectivement aux sondes 97 et 77. Pourtant, ces sondes ne sont pas issues de gènes ayant des relations (analyse avec Synteny Search). Il est possible que les sondes ne soient pas très spécifique ou qu'il y ait eu un artéfact avec les clones BAC ou l'hybridation.

Au final, 16 sondes présentent en couple au moins 1 clone BAC commun et 3 groupes de sondes, constitués de respectivement 4, 6 et 7 sondes, présentent plusieurs clones BAC communs avec des relations complexes. Malgré des clones BAC communs à 2 ou plusieurs sondes, aucune des sondes ne sont proches chez *Arabidopsis*.

Afin de vérifier l'efficacité des sondes Overgo, des amplifications PCR ont été faites sur les clones BAC positifs à 10 sondes Overgo (sondes 17, 22, 27, 39, 97, 103, 108 114, 79 et 86) à l'aide de 10 couples d'amorces (P05 à P14, pour plus de détails sur les couples d'amorces, voir page 127). La figure 3.10 présente les résultats de l'amplification avec ces amorces. Les 5 premiers gels (P05 à P9) correspondent aux dépôts de la légende en haut à gauche, les quatre derniers gels (P11 à P14) correspondent aux dépôts au milieu à gauche. Pour le gel P10, le dépôt est différent des autres mais seul le clone BAC 373D09 est amplifié par les amorces P10. Toutes les amorces utilisées fonctionnent et amplifient au moins un clone BAC. Les amorces utilisées pour l'amplification ont été définies à partir de la même séquence utilisée pour définir les sondes Overgo. Ainsi la sonde Overgo 17 a été définie à partir de la même séquence que les amorces P05, de même pour la sonde 22 et P06, la sonde 27 et P07, la sonde 39 et P08, la sonde 97 et P09, la sonde 103 et P10, la sonde 108 et P11, la sonde 114 et P12, la sonde 79 et P13 et enfin la sonde 86 et P14.

Les amorces P07, P10, P11 et P14 amplifient une bande pour chacun des clones BAC spécifiques des sondes Overgo correspondantes (exemple : P07 n'amplifie que le clone 274F11 spécifique de la sonde 27). Les amorces P08, P12 et P13 n'amplifient que certains clones BAC positifs des sondes Overgo correspondantes. Ainsi, P12 n'amplifie que le clone BAC 146H04 alors que 4 clones étaient positifs avec la sonde 79. De même, P13 n'amplifie que le clone 73C20 alors que 276C10 était aussi positif avec la sonde 79. Sur les 19 clones BAC spécifiques de la sonde Overgo 39, 8 ne sont pas amplifiés par les amorces P08. Quant aux amorces P05, P06 et P09, elles amplifient les clones BAC positifs aux sondes correspondantes mais aussi d'autres clones BAC. Les amorces P06 amplifient le clone BAC 12F04 ainsi que le clone BAC 34E04. Les amorces P05 amplifient en plus des clones BAC spécifiques le clone BAC 274F11. Les amorces P09 amplifient en plus des clones BAC spécifiques 6 autres clones BAC. La grande majorité des clones identifiés par des sondes Overgo sont amplifiés à l'aide des amorces définies par Iccare permettant ainsi de vérifier l'efficacité des sondes Overgo.

Les amorces P05 et P08 amplifient différentes tailles de fragments en fonction du clone BAC. Ainsi, l'amplification des amorces P05 permet de distinguer les clones 103O23 et 406L11 des clones 272K13, 269E13, 330E09 et 274F11. Les amorces P08 permettent d'amplifier 11 clones BAC sur les 15 spécifiques de la sonde 39. Ces 11 clones sont différenciés par la taille du fragment amplifié. Six clones BAC ont une bande "haute", 4 clones BAC ont une bande "basse" et un clone BAC possède les deux bandes. Les sondes Overgo définies ne sont pas spécifique d'un gène mais de l'ensemble des gènes qui présentent des similitudes pour la région dans laquelle la sonde Overgo a été définie.

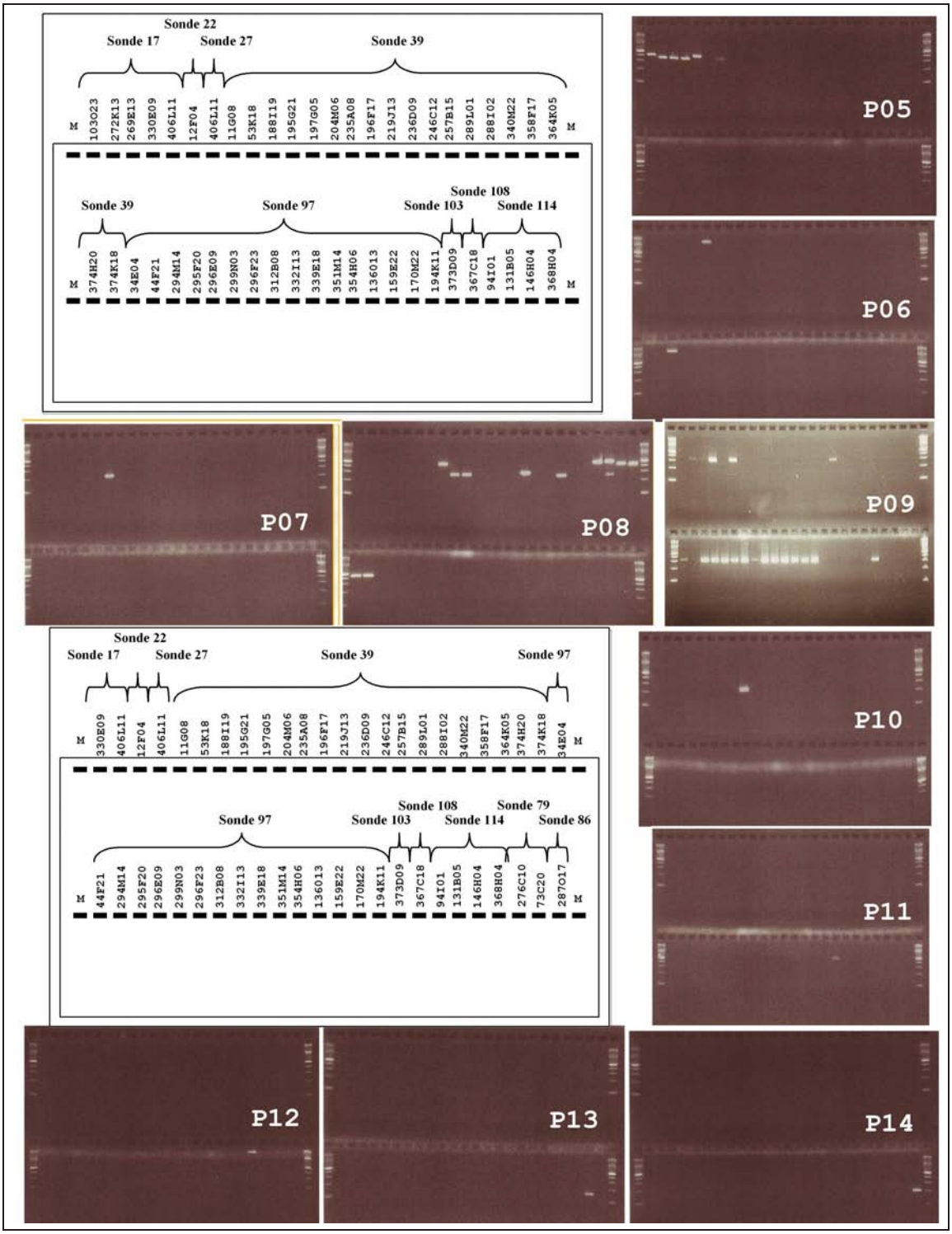


FIG. 3.10 – Amplification PCR sur les clones BAC positifs à 10 sondes Overgo.

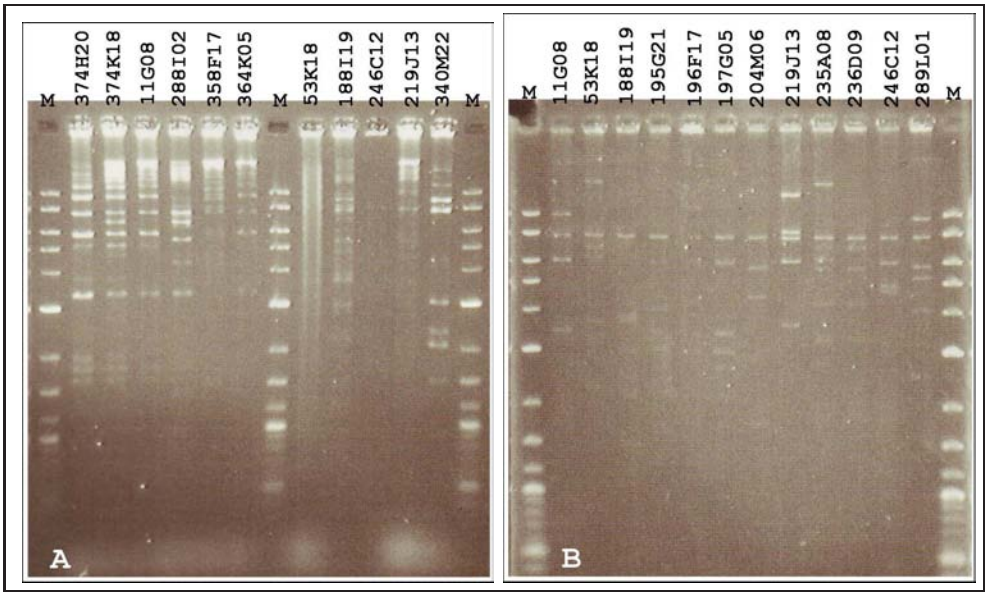


FIG. 3.11 – Profil de digestion *Hind*III de clones BAC spécifiques de la sonde 39.

Nous n'avons pas réussi à regrouper de clones BAC en utilisant la technique des sondes Overgo communes, nous avons donc extraits les ADN de quelques clones BAC positifs à des sondes proches chez *Arabidopsis* que nous avons digérés par l'enzyme *Hind*III puis séparé les produits de digestion sur gel d'agarose 1% afin de créer une "empreinte de digestion" (fingerprint) pour regrouper les clones qui seraient chevauchant. Parallèlement, nous avons aussi digéré l'ADN des clones BAC positifs de la sonde 39 qui ont pu être regroupés à l'aide des amplifications PCR avec le couple d'amorces P08.

La migration des produits de digestion des clones BAC spécifiques de la sonde 39 est présentée Figure 3.11. Les migrations ne sont pas de très bonne qualité et ne sont que des résultats préliminaires. Sur le gel A, les clones 53K18 (2), 188I19 (3), 219J13 (8), 246C12 (11) et 358F17 (15) n'ont pas correctement été digérés. Sur le gel B, les clones 188I19 (3), 195G21 (4) et 196F17 (5) sont trop faible pour être analysés. Les produits de digestion du clone 11G08 (1) et 374H20 (17) présentent les mêmes 5 bandes (sans compter la bande commune à tous les clones BAC et correspondant au vecteur). Le clone 374K18 (18) présente les 5 mêmes bandes que les clones 11G08 et 374H20 mais aussi 2 autres bandes qui elles sont identiques à celles du clone 288I02 (13). Le clone 364K05 (16) présente 3 bandes communes aux clones 11G08, 374H20 et 374K18. Les autres clones ne semblent pas présenter de bandes communes. Cinq des clones BAC identifiés par la sonde 39 présentent des bandes communes et seraient donc chevauchant. Tous les clones BAC ne semblent pas chevauchant ce qui laissent à penser qu'il y aurait plus d'un gène spécifique de la sonde 39 et donc plusieurs localisations chromosomiques.

Les clones BAC spécifiques des sondes 141, 114, 71, 115, 72, 116, 142 et 143 ont aussi été digérés et séparés sur gel d'agarose mais aucun des clones BAC ne présente de bande commune (résultats non présentés). Les clones BAC spécifiques des sondes 141, 114, 71, 115, 72, 116, 142 et 143 ne peuvent être regroupés en clones chevauchants.

3.3 Discussion des Résultats Expérimentaux

3.3.1 Iccare permet de définir des Marqueurs Moléculaires Exploitables

Les séquences EST et ARNm utilisés pour construire la carte génétique de Gentzbittel *et al.* de 1999 ont été comparées aux gènes d'*Arabidopsis*. Sur l'ensemble des séquences utilisées seul 34% de celles-ci présentent des similitudes avec des gènes d'*Arabidopsis*. Ce pourcentage est identique à celui obtenu lorsque l'on compare l'ensemble des séquences EST et ARNm des bases de données publiques (séquences soumises à Iccare, soit 60 200 séquences). Seul 34% des séquences de tournesol présentent des similitudes avec des gènes d'*Arabidopsis*.

Pourtant, entre *Arabidopsis* et le riz, le nombre de gènes qui sont homologues est à peu près de 80% (Goff *et al.*, 2002, Yu *et al.*, 2002). Le tournesol est plus proche taxonomiquement d'*Arabidopsis* que le riz il est donc très fortement probable que le taux d'homologie entre les gènes du tournesol et d'*Arabidopsis* soit équivalent ou supérieur à 80%. Hors, sur l'ensemble des séquences de tournesol utilisées, le pourcentage de séquences similaires n'est que de 34%. Ce pourcentage est plus faible que le taux d'homologie entre les gènes d'*Arabidopsis* et du riz et ceci est principalement dû à la nature des séquences EST de tournesol.

La grande majorité des séquences EST de tournesol sont des séquences en 5' ou 3', c'est-à-dire correspondant aux régions non traduites des ARNm. Un ARNm est constitué d'une région 5' non traduite ou UTR 5' (UnTranslated Region 5'), d'une région codante (traduite pour donner la protéine) et d'une partie 3' non traduite ou UTR 3' (UnTranslated Region 3'). Les régions UTR sont très polymorphes et présentent donc peu de similitude lorsque l'on compare des gènes homologues. En revanche, les régions traduites sont plus conservées lorsqu'on les compare entre elles. Les séquences EST de tournesol utilisées sont majoritairement des séquences 5' et 3'. Il est donc normal qu'un certain nombre de séquences EST en 5' ou 3' ne présentent pas de similitudes avec des gènes d'*Arabidopsis* ; cependant cela ne signifie pas que ces séquences n'ont pas de gènes homologues chez *Arabidopsis* mais seulement que la partie de séquence dont on dispose ne correspond qu'à la partie la moins conservée de la séquence.

Afin d'augmenter le pourcentage de similitude entre les séquences de tournesol et les gènes d'*Arabidopsis*, la comparaison doit se faire avec des séquences complètes, ARNm pleine longueur ou full length mRNA. La comparaison des séquences EST de tournesol avec les gènes d'*Arabidopsis* ne donnent que 34% de séquences similaires alors qu'avec les séquences ARNm de tournesol le pourcentage de similitudes montent à 60%. Avec des séquences complètes, ARNm pleine longueur ou full length mRNA, les résultats de comparaison de séquences entre tournesol et *Arabidopsis* donnent de bien meilleurs résultats mais le nombre d'ARNm pleine longueur disponible pour le tournesol est actuellement très faible 356 séquences soit à peine 1% des gènes présents chez la plante modèle *Arabidopsis*.

A partir des séquences de tournesol qui présentent des similitudes avec des gènes d'*Arabidopsis*, nous avons défini des couples d'amorces dans des exons contigus afin d'amplifier les régions introniques de l'ADN génomique du tournesol. Les 51 couples d'amorces ont été définis à partir de séquences provenant de différents cultivars de tournesol (Ha280, Ha801, Emil, psc8) et sont testés sur différents cultivars de tournesol (PAC2, RHA266, AS613, SD1', psc8). Au final, 5 couples d'amorces n'amplifient aucun cultivar (39, 48, 49, 50 et 51), 3 couples d'amorces amplifient faiblement quelques cultivars (16, 28 et 36) et 43 couples d'amorces amplifient tous les cultivars. Pour les amorces qui n'amplifient rien, il est possible que la taille de l'intron soit trop importante, qu'il y ait eu des problèmes avec le mixte PCR ou que les conditions PCR ne soient pas optimales. Personnellement, je pencherai plutôt pour les conditions de PCR, car le couple d'amorces 28, qui fonctionne faiblement, a été défini à partir de la séquence de rubisco de part et d'autre de l'intron 2 et on sait que ces amorces fonctionnent, c'est juste un problème d'ajustement de condition PCR. Pour vérifier les résultats il suffira de reproduire les amplifications PCR en modifiant les conditions. Toujours est-il qu'au final, quelle que soit l'origine de la séquence utilisée, les amorces qui fonctionnent sur un cultivar fonctionnent aussi sur les autres cultivars et les amorces définies à l'aide d'Iccare sont efficaces à plus de 90%, la seul limitation est la taille de l'intron (doit être inférieure à 1 000 pb).

A partir des couples d'amorces définis, nous avons chercher à obtenir de nouveaux marqueurs moléculaires afin de construire ou d'incorporer de nouveaux marqueurs moléculaires au sein des cartes génétiques. Les produits d'amplification doivent donc présentés du polymorphisme entre cultivars et surtout entre les lignées parentales qui ont servi à la construction des cartes génétiques.

Certains couples d'amorces présentent effectivement du polymorphisme de taille visible sur agarose. Cette différence de polymorphisme de taille nous a permis de différentier le cultivar RHA266 des cultivars AS613, SD1', psc8 et PAC2 par la taille des fragments amplifiés. Les résultats de migration des produits d'amplification de la rubisco (voir page 70) en SSCP ont montré que psc8 et PAC2 ont des profils de migration identiques alors que RHA266 a un profil de migration identique à Ha821, Ha300 et KA. A priori, les cultivars AS613 et SD1' se rapprocheraient plus de psc8 et PAC2. Le polymorphisme de taille nous permet de différentier les cultivars de tournesol entre eux et il est possible d'établir des relations de phylogénie entre ces cultivars de tournesol.

La différence de taille des produits d'amplification entre PAC2 et RHA266 (parents des lignées recombinantes) va nous permettre de transformer nos couples d'amorces en marqueurs moléculaires. Parmi les 19 couples d'amorces qui semblent présenter du polymorphisme de taille sur gel d'agarose 2%, 8 ont été testés sur gel agarose 4%. Sept des 8 couples d'amorces présentent effectivement du polymorphisme de taille visible sur agarose entre les lignées PAC2 et RHA266. Nous avons donc au final pour les 51 couples d'amorces définies une efficacité d'amplification de 90% et un pourcentage de marqueurs moléculaires vérifiés de 17% (calculé par rapport aux amorces qui fonctionnent). Sept couples d'amorces sur 8 présentent du polymorphisme pour les amorces testées soit environ 90%, avec 19 couples d'amorces potentiellement polymorphes, on pourrait atteindre environ 35% de marqueurs moléculaires polymorphes sur

agarose. Ces résultats sont confirmés par ceux obtenus par le GénoPlante qui ont exploité Iccare pour définir environ 400 couples d'amorces qui présentent 30% de polymorphisme de taille sur gel d'agarose et plus de 65% de polymorphisme de taille sur séquenceur automatique (Delphine Samson, communication personnelle).

Sept marqueurs moléculaires sur les 8 testés pour leur polymorphisme entre RHA266 et PAC2 ont été intégrés à la carte génétique construite par Rachid Al-Chaarani *et al.* (2004). Un LOD score de 4.0 a été utilisé ce qui nous a permis d'assigner 4 marqueurs moléculaires à 4 groupes de liaisons différents et 3 marqueurs sont non liés à la carte génétique. Ces marqueurs moléculaires permettent de venir compléter les données déjà obtenues sur la carte génétique. Quatre marqueurs sur 7 ont pu être cartographiés ce qui illustre bien le fait que la carte n'est pas encore saturée, et que certaines régions du génome ne sont pas représentées. Ces marqueurs moléculaires sont d'une grande importance au sein des cartes génétiques car ils sont associés à des gènes exprimés alors que la majorité des marqueurs des cartes génétiques sont construites à partir de marqueurs anonymes. L'utilisation de couples d'amorces qui amplifient les régions introniques nous permet de mettre au point des marqueurs moléculaires rapidement identifiables et cartographiables pour augmenter la saturation des cartes génétiques.

3.3.2 Iccare permet de définir des Sondes Overgo Efficaces

Les sondes Overgo ont été définies dans les régions conservées des séquences EST ou ARNm de tournesol puis hybridées par pool sur l'ensemble de la banque de clones BAC. Ces sondes en pool ont permis de récupérer 503 clones BAC positifs (clones BAC récupérés) et 330 clones BAC ont été associés sonde par sonde aux 161 sondes Overgo utilisées (clones BAC identifiés). Le résultat du criblage d'une banque de clones BAC dépend de trois facteurs : la qualité de la banque, la qualité des membranes et les sondes utilisées.

La qualité des membranes de la banque de clones BAC a été vérifiée en comparant la différence entre clones BAC récupérés (hybridation en pool) et identifiés (hybridation individuelle) pour chacune des membranes. Cette différence est principalement due à l'utilisation en pool des sondes pour la récupération des clones BAC positifs lors du criblage des membranes. Les hybridations ont été réalisées pour chaque sonde deux fois sur deux lots différents de membranes. Nous avons récupéré tous les clones BAC positifs qui apparaissaient au moins une fois lors des hybridations. Un certain nombre de faux positifs a donc été récupéré. Nous avons alors étudié la différence entre clones BAC récupérés et identifiés pour chaque membrane de la banque de clones BAC et il apparaît qu'il existe un effet membrane qui influence le pourcentage de clones BAC récupérés/identifiés. Ainsi, la membrane B présente un pourcentage de clones BAC récupérés/identifiés de 43% alors que pour les autres membranes il est supérieur à 63% (et peut atteindre 85%). Cette différence est due à la qualité de la membrane.

Les membranes B sont les premières membranes à avoir été dupliquées et repiquées. Elles ont eu un temps de pousse différent des autres membranes qui ont été repiquées par la suite

(environ 3 à 4 heures en moins de pousse). Les clones BAC déposés sur les membranes B contiennent moins d'ADN. Les membranes ont toutes subi les mêmes rinçages après l'hybridation ainsi que le même temps d'exposition. Les membranes B ayant moins de matériel génomique ont été trop rincées et pas assez exposées en comparaison des autres membranes. La lecture de ces membranes B a donc été plus difficile que pour les autres membranes et un certain nombre de clones qui semblaient positifs ont été récupérés (et par principe de précaution, mieux vaut en récupérer plus que pas assez). Avec des rinçages moins importants et une exposition plus longue, on devrait pouvoir obtenir des lectures des membranes B plus propres et le pourcentage de récupération/identification devrait, de ce fait, augmenter et se rapprocher de ceux observés pour les autres membranes.

La qualité des membranes est donc un facteur important dans l'interprétation des résultats de l'hybridation : plus la qualité des membranes est élevée, plus l'adaptation du protocole de rinçage et d'exposition sera facilitée et meilleur sera la lecture des membranes. Si on ne tient pas compte de la membrane B, les résultats des autres membranes montrent que l'utilisation des sondes en pool permet de récupérer un nombre de clones BAC qui sera identifié par la suite à plus de 65% (soit plus de deux clones sur trois).

Un autre facteur intervenant dans les résultats de l'hybridation correspond aux sondes utilisées. Les sondes Overgo que nous avons utilisées ont été définies selon plusieurs critères : l'organisme, le cultivar, les similitudes avec les gènes d'*Arabidopsis*. Les séquences utilisées pour la définition des sondes proviennent principalement du tournesol (150 séquences) mais 11 séquences proviennent d'autres organismes qui présentaient des similitudes avec au moins 4 organismes dont *Arabidopsis* (10 sondes Overgo faites à partir de séquences de *Medicago truncatula* et 1 à partir d'une séquence de maïs). Les sondes Overgo faites à partir de séquences de tournesol présentent un taux d'efficacité (s'hybrident au moins avec un clone BAC) d'environ 70% (sondes Overgo e^{-20} et e^{-5}) alors que celles faites à partir de séquences d'un autre organisme ne sont efficaces qu'à 25%. Cette différence d'efficacité est principalement due au fait que les sondes Overgo ont une taille de 40 nucléotides et sont donc très spécifiques de la région génomique correspondante. Les nucléotides polymorphes entre les séquences des organismes qui ont servi de matrice avec la séquence génomique du tournesol peuvent être suffisants pour que la sonde ne s'hybride pas. Il est donc préférable de travailler à partir de séquences appartenant à l'organisme étudié.

Le deuxième critère qui a été étudié est le cultivar utilisé pour définir les sondes. La banque de clones BAC a été construite à partir du cultivar Ha821, les sondes Overgo ont été définies à partir de séquences provenant de différents cultivars (Emil, Ha280, Ha801, psc8 et d'autres indéfinis [séquence TC, ARNm]). Le résultat des hybridations avec des sondes Overgo issues de cultivars différents de celui de la banque de clones BAC n'est pas fonction du cultivar de la séquence utilisé (résultats non présentés). Il n'y a pas de différence significative entre les résultats des sondes Overgo définies à partir de séquences Ha280, Ha801, psc8 ou Emil. Il est donc possible de travailler à partir de n'importe quel cultivar de l'espèce étudié pour définir des sondes Overgo.

Le dernier critère étudié correspond aux similitudes avec les gènes d'*Arabidopsis*. Les similitudes entre les séquences de tournesol et les gènes d'*Arabidopsis* ont un impact sur l'efficacité

des sondes Overgo. Deux types de sondes Overgo ont été définies à partir de séquences de tournesol. Des sondes Overgo e^{-20} définies à partir de séquences de tournesol de *fortes similitudes* et des sondes Overgo e^{-5} définies à partir de séquences de tournesol de *faibles similitudes*. L'efficacité des sondes Overgo e^{-20} est de 72% alors que celle des sondes Overgo e^{-5} est de 61%. Malgré le fait que les sondes Overgo sont définies dans la région qui présentent des similitudes entre les séquences de tournesol et d'*Arabidopsis* (que ce soient les Overgo e^{-20} ou les e^{-5}), les sondes Overgo e^{-5} ont une efficacité un peu moindre. Cette différence d'efficacité peut s'expliquer par le fait que les séquences de *faibles similitudes* présentent un nombre de nucléotides polymorphes un peu plus important que les séquences de *fortes similitudes*. Ces nucléotides polymorphes ont un impact plus important sur une sonde Overgo de 40 nucléotides que sur une une sonde plus longue (type PCR). Il est donc normal que ces nucléotides polymorphes influs sur l'efficacité des sondes Overgo. Il est donc préférable de travailler à partir de séquences qui présentent de fortes similitudes avec les gènes d'*Arabidopsis* afin d'optimiser l'efficacité des sondes.

Le dernier facteur qui influence le résultat des hybridations correspond à la couverture de la banque utilisée. La banque de clones BAC utilisée est constituée de 156 672 clones BAC contenant chacun un fragment d'ADN génomique de tournesol d'une taille moyenne de 80 Kbases. Cette banque de clones BAC est donc équivalente à environ 4 à 5 fois le génome du tournesol (banque 4 à 5 fois couvrante), ce qui signifie que pour une sonde unique dans tout le génome nous devrions récupérer 4 à 5 clones BAC. Nous avons donc regardé le nombre de clones BAC positifs pour chaque sonde utilisée. Le nombre de clones BAC positifs divisé par le nombre de sondes qui se sont hybridées donne une moyenne de 3,56 clones BAC positifs par sonde (en ne tenant pas compte des sondes qui ne se sont pas hybridées). Cependant, la distribution du nombre de clones BAC par sonde montre que 72% des sondes qui s'hybrident présentent de 1 à 3 clones BAC positifs (56% de 1 à 2 clones BAC). Ces résultats laissent à penser que la banques de clones BAC seraient plutôt 2 à 3 fois couvrantes pour la région étudiées. Avec ce taux de couverture il est pas impossible que certaines régions ne soient pas représentées au sein de la banque de clones BAC ce qui implique qu'un certain nombre de sondes Overgo ne sont pas hybridées non pas à cause d'un manque d'efficacité mais de l'absence de la région génomique correspondante. L'efficacité des sondes Overgo est donc sous estimée avec cette banque de clones BAC. La qualité de la banque de clones BAC utilisée est donc primordiale pour de bonne hybridation.

Nous nous sommes ensuite intéressé aux sondes Overgo qui s'hybrident avec beaucoup de clones BAC (supérieur à 10 clones BAC) alors que la majorité des sondes ne s'hybrident qu'avec 2 ou 3 clones BAC. C'est le cas, par exemple, de la sonde r8 spécifique de la rubisco qui s'hybride avec 20 clones BAC, de la sonde 97 spécifique d'une protéine kinase calcium dépendante qui s'hybride avec 16 clones BAC ou encore de la sonde 39 spécifique d'une protéine Histone H3 qui s'hybride avec 19 clones BAC. Mais, en regardant de plus près les gènes utilisés pour définir ces sondes, il apparaît qu'ils appartiennent à des familles multigéniques.

La séquence du gène de rubisco At5g38430 similaire à la séquence EST de tournesol qui a servi à définir la sonde r8 est similaire à 3 autres gènes chez *Arabidopsis*, et chez le tournesol, on sait que la rubisco est codée par au moins 5 gènes (voir page 80). Seize clones BAC ont été récupérés et au moins 5 gènes existent chez le tournesol ce qui nous donne 3 clones BAC par gènes.

La séquence du gène de la protéine Histone H3 At5g10980 similaire à la séquence EST de tournesol qui a servi à définir la sonde 39 est similaire à 8 autres gènes chez *Arabidopsis* mais seulement 4 sont similaires dans la région qui a servi à définir la sonde. Chez le tournesol, le nombre de gènes présents n'est pas connu, mais des amplifiations sur plusieurs cultivars de tournesol avec les amorces P08 (définies à partir de la même séquence qui a servi à définir la sonde) ont montré un profil d'amplification complexe ayant 2 ou 3 bandes de tailles différentes sur gel d'agarose. Les clones BAC positifs de cette sonde ont été amplifiés par ces amorces P08 et les produits d'amplification présentent une bande unique de taille différente en fonction du clone BAC. Les différents clones BAC permettent ainsi de reconstituer le profil d'amplification des cultivars. Ces clones BAC ont été digérés et un fingerprint a été obtenu pour plusieurs d'entre eux ce qui a permis de regrouper plusieurs clones BAC qui présentaient un fragment amplifié par P08 de même taille ($11G08$, $374H20$, $374K18$ et $346K05$). Il semblerait qu'il y ait au moins 3 gènes différents chez le tournesol. Bien entendu, il faudrait séquencer les différents clones BAC afin de vérifier qu'ils sont bien recouvrants mais aussi pour déterminer le nombre exact de gènes.

La séquence du gène de la protéine kinase calcium dépendante At5g12480 similaire à la séquence EST de tournesol qui a servi à définir la sonde 97 est similaire à 12 autres gènes d'*Arabidopsis* mais seulement 6 sont similaires dans la région qui a servi à définir la sonde. En revanche, le nombre de gènes chez le tournesol est inconnu (une seule bande visible sur agarose, mais tout comme la rubisco cette bande peut cacher plusieurs gènes).

Ces 3 exemples permettent de mettre en avant que toutes les sondes ne sont pas unique et ceci explique le nombre plus important de clones BAC positifs pour certaines sondes.

Afin de tenir compte du fait que certaines sondes ne sont pas uniques à un seul gène, la moyenne et la distribution du nombre de clones BAC par sonde va être pondérée par le nombre de gènes potentiels pour chacune des sondes. Le nombre de clones BAC positifs a donc été divisé non pas par le nombre de sondes qui s'hybrident mais par le nombre de gènes d'*Arabidopsis* similaires aux gènes d'*Arabidopsis* qui sont similaires aux séquences de tournesol (même si le nombre de gènes entre *Arabidopsis* et le tournesol doit certainement varier). La nouvelle moyenne calculée est de 1,8 et la distribution montre que 84% des gènes potentiels présentent de 1 à 3 clones BAC positifs (63% de 1 à 2 clones BAC). Il semblerait que la banque de clones BAC serait plutôt 1 à 2 fois couvrantes. D'autres indices indiques que la banque de clones BAC n'est pas très couvrante dans la région étudiée. Certaines sondes Overgo définies à partir de gènes proches chez *Arabidopsis* ne s'hybrident pas. C'est le cas des sondes 98, 48, 49 et 99 ou des sondes 87, 88, 126, 89 et 127 qui pourtant sont toutes définies à partir de séquences EST de tournesol et devraient de ce fait fonctionner. De plus, les amorces P03 et P15 définies à partir de séquences EST qui ont servi à définir les sondes Overgo 12 et 132

amplifient les différents cultivars de tournesol alors que les sondes Overgo ne s'hybrident avec aucun clone BAC (hybridation réalisée deux ou trois fois sur des membranes différentes).

Au regard de ces résultats, il semble que la banque de clones BAC ne soit pas 4 à 5 fois couvrantes dans la ou les régions étudiées ; elle serait plutôt 2 fois couvrantes et présenterait aussi des "gaps" (certaines régions ne sont pas représentées). Puisque la banque de clones BAC ne semblent pas aussi couvrantes et surtout semble présenté des gaps pour la ou les régions étudiées, le nombre de sondes Overgo qui s'hybrident doit aussi être légèrement sous estimé. Les pourcentage d'efficacité des sondes Overgo e^{-20} et e^{-5} sont donc des valeurs minimum d'efficacité. L'efficacité des sondes Overgo e^{-20} est de 72% minimum et l'efficacité des sondes Overgo e^{-5} est de 60% minimum. Ces résultats sont d'ailleurs confirmés par ceux obtenus par d'autres équipes aussi bien chez les animaux que chez les végétaux. Ainsi, Gardiner *et al.* (2004) ont utilisé 10 642 sondes Overgo pour cribler une banque de clones BAC de maïs. Environ 88% des sondes Overgo s'hybrident avec au moins un clone BAC avec une moyenne d'environ 10,7 clones BAC par sonde (sans comptabiliser les sondes qui s'hybridaient avec plus de 25 clones BAC (sondes spécifiques de gènes appartenant à des familles multigéniques)). L'efficacité des sondes Overgo chez les animaux (Cai *et al.*, 1998, Han *et al.*, 2000) est respectivement de 92 et 91%. La différence entre les résultats des autres équipes et les nôtres (environ 15 à 20% d'efficacité de moins pour les sondes Overgo définies à partir des séquences de tournesol de *fortes similitudes*) semblent confirmer que la banque de clones BAC du tournesol n'est pas suffisamment couvrante dans la ou les régions étudiées.

3.3.3 Organisation du Génome du Tournesol et Synténie

Les informations fournies par l'analyse de la banque de clones BAC et des marqueurs moléculaires cartographiés nous ont permis d'obtenir des informations sur l'organisation du génome du tournesol, ou tout du moins une partie du génome. Dans un premier temps nous nous sommes intéressés au nombre de gènes que contient le génome du tournesol.

Les sondes Overgo utilisées ont montré que le nombre de clones BAC positifs à une sonde est corrélé au nombre de gènes similaires et spécifiques de cette sonde ; c'est-à-dire que les gènes d'*Arabidopsis* qui appartiennent à une famille multigénique le sont également chez le tournesol, de même pour les gènes qui sont uniques chez l'un le sont généralement chez l'autre. De même, les séquences EST de tournesol de la carte génétique F2 qui sont localisés à plusieurs endroits dans le génome du tournesol sont similaires à des gènes d'*Arabidopsis* qui sont dupliqués ou similaires à au moins un autre gène d'*Arabidopsis*. Les clones BAC positifs de certaines sondes (17, 27, 39 et 132) amplifiées par les amorces P5, P7, P8 et P15 permettent de mettre en évidence la présence de plusieurs gènes pour ces amorces, conformément aux résultats indiqués par Synteny Search pour les gènes d'*Arabidopsis*. Les amplifications sur les différents cultivars de tournesol avec les amorces P4, P29, P40, P43 et P46 amplifient aussi plusieurs gènes. Les autres couples d'amorces ne présentent qu'une seule bande sur agarose mais ils peuvent aussi amplifier plusieurs gènes, comme pour la rubisco qui ne présente qu'une bande mais qui amplifie plusieurs gènes (au moins 5). Le tournesol semble même avoir des gènes dupliqués qui ne le sont pas chez *Arabidopsis*. Par exemple, les amorces P05 amplifient

deux bandes alors que le *gène similaire* d'*Arabidopsis* est unique. Certaines séquences EST de la carte génétique des F2 (marqueurs S066, S301 et S098) qui sont localisées à deux loci sont similaires à des gènes uniques chez *Arabidopsis*.

Même si le nombre de gènes similaires chez *Arabidopsis* n'est pas identique à celui des gènes similaires chez le tournesol, la corrélation entre les deux est tout de même forte. A partir de ces données nous avons estimé le nombre de gènes dans le génome du tournesol et nous estimons que le génome du tournesol contient environ 1,5 à 2 fois le nombre de gènes présents chez *Arabidopsis*, soit entre 40 000 et 50 000 gènes. Ce nombre de gènes rapprocherait le tournesol du riz qui en possèderait environ 40 000 (riz : génome de 450 Mbase, 2x=2n=12) ou du peuplier qui en disposerait d'autant (peuplier : génome de 550 Mbase, 2x=2n=19).

L'hypothèse de travail de départ pour étudier l'organisation des gènes au sein du génome du tournesol était qu'en hybridant des sondes Overgo définies à partir de séquences de tournesol similaires à des gènes distants d'environ 45 Kbases chez *Arabidopsis* nous récupérions dans la banque des clones BAC qui sont recouvrants, nous permettant ainsi de construire une carte physique. La taille moyenne des inserts des clones BAC est de 80 Kbases, donc notre hypothèse nous prédisait environ 2 sondes Overgo par clone BAC.

Les sondes Overgo nous ont effectivement permis d'identifier un certain nombre de clones BAC, mais très peu d'entre eux sont communs à plusieurs sondes. La couverture de la banque de clones BAC dans la région étudiée est seulement une à deux fois couvrante, ce qui ne nous a pas permis de récupérer suffisament de clones BAC pour chercher des recouvrements. Les quelques sondes qui présentent des clones BAC en commun, comme les sondes 94 et 95 qui s'hybrident avec les mêmes 4 clones BAC ou les sondes 7 et 136 qui s'hybrident avec les mêmes 3 clones BAC, ne nous ont pas permis de constuire une carte physique. Les sondes 94 et 95 sont normalement suffisamment proches (en tout cas chez *Arabidopsis*) pour présenter des clones BAC communs et c'est le cas, sauf que ces sondes ont été faites à partir de deux gènes qui sont similaires (duplication en tandem chez *Arabidopsis*). Les sondes 7 et 136 ne sont pas proches chez *Arabidopsis*, en revanche, les gènes qui ont servi de matrice sont similaires, il est donc normal d'avoir des clones BAC communs entre ces 2 sondes mais ces clones BAC ne doivent pas être recouvrant.

Ces résultats nous ont donc permis d'émettre 2 hypothèses : la synténie est conservée entre le génome d'*Arabidopsis* et celui du tournesol mais la distance entre les gènes du tournesol est plus importante que chez *Arabidopsis* ou la synténie entre le génome d'*Arabidopsis* et celui du tournesol n'est pas conservée.

En partant de l'hypothèse que la synténie est conservée entre le génome du tournesol et celui d'*Arabidopsis*, l'espacement entre les gènes de tournesol serait trop grand pour avoir des sondes communes à un ou plusieurs clones BAC. L'espacement entre les gènes chez *Arabidopsis* est d'environ 5 Kbases, chez le tournesol si on considère qu'il y a entre 40 00 et 50 000 gènes sachant que la taille du génome est de 3 000 Mbases alors on aurait environ 1 gène tous les 60 - 75 Kbases soit environ 1 gène par clone BAC. Les fingerprint effectués sur des clones BAC contenant des gènes sensés être proche (chez *Arabidopsis*) n'a pas permis de faire des

regroupement de clones BAC. Mais il est normal de ne pas pouvoir regrouper les clones BAC puisque la banque de clones BAC est peu couvrante dans cette région et que l'espace entre les gènes est d'environ la taille d'un clone BAC.

Inversement, si la synténie entre *Arabidopsis* et le tournesol n'est pas conservée, ce qui peut aussi expliquer nos résultats, alors la répartition des gènes, qui étaient proche chez *Arabidopsis*, est aléatoire au sein du génome du tournesol ce qui explique pourquoi les sondes Overgo ne présentent pas de clones BAC communs et qu'il est impossible de faire des regroupements de clones BAC.

Les résultats de ce travail de thèse sont actuellement insuffisants pour valider ou exclure l'une des deux hypothèses.

La comparaison entre le tournesol et *Arabidopsis* nous a permis d'étudier le génome du tournesol et d'essayer de déterminer la conservation de synténie entre le génome de ces 2 plantes. L'utilisation *Arabidopsis* comme plante modèle nous a permis de transposer les informations structurelles des gènes d'*Arabidopsis* aux séquences du tournesol. Ce transfert d'information (position des introns principalement) est efficace et nous a permis de mieux exploiter les séquences de tournesol en définissant des couples d'amorces de part et d'autre des introns ou ds sondes Overgo dans les exons conservés des gènes.

Pourtant, il n'a pas été possible de déterminer si la synténie entre la plante modèle et le tournesol est conservée ou non, aucune de nos données ne permet de la confirmer ou de l'infirmer. Les séquences EST correspondant aux marqueurs moléculaires de la carte F2 (Gentzbittel *et al.*, 1999) ont été comparées aux séquences codantes des gènes d'*Arabidopsis* mais seul 41 séquences présentent des similitudes avec les gènes d'*Arabidopsis*. Ces séquences similaires ne sont pas assez nombreuses pour définir des blocs de conservation entre chromosomes de tournesol et chromosomes d'*Arabidopsis*. De même, avec les 7 marqueurs moléculaires cartographiés sur la carte des RILs (Rachid Al-Chaarani *et al.*, 2004), ces données ne permettent pas de déterminer si des blocs macro-synténiques entre le tournesol et *Arabidopsis* existent.

La micro-synténie entre le génome du tournesol et *Arabidopsis* n'a pas pu être établie. Les sondes Overgo utilisées nous a permis d'identifier des clones BAC contenant les gènes mais aucun des clones BAC n'a pu être regroupés au sein de contig afin de construire une carte physique. Aucune de nos données ne permis confirmer ou d'exclure l'hypothèse de conservation de synténie entre le génome du tournesol et celui d'*Arabidopsis*.

La synténie a déjà été établie entre des espèces relativement éloignées phylogéniquement, comme entre *Arabidopsis* et le riz ou entre *Arabidopsis* et *Medicago*. Le tournesol devrait, théoriquement, aussi montré de la synténie avec le génome d'*Arabidopsis* mais le problème est de pouvoir déterminer quel est le degré de conservation de synténie qu'il existe entre ces 2 plantes. Entre plantes proches, *Arabidopsis* et les Brassica, on observe quelques remaniements chromosomiques et les blocs synténiques entre les génomes ont des tailles de quelques Mbases. Avec *Medicago*, une plante plus éloignée que les Brassica vis-à-vis d'*Arabidopsis*, la synténie a aussi été établie, mais les remaniements de blocs chromosomiques sont plus fréquents et leur

taille est moindre (plusieurs Kbases). C'est la même chose avec la synténie entre *Arabidopsis* et le riz (2 plantes très éloignées, car la première est une dicotylédone et l'autre est une monocotylédone). Les blocs chromosomiques synténiques sont encore plus petits, quelques Kbases, et les remaniements sont très fréquents.

La même chose est observée au niveau de la micro-synténie. Entre plantes proches, la position et l'ordre des gènes contenus dans un bloc chromosomique synténique vont être très fortement conservés, et plus on va comparer des plantes éloignées plus on va observer la perte ou l'insertion de gènes et des duplications en tandem au sein des gènes conservées du bloc chromosomique synténique. Il est tout à fait possible qu'entre le génome du tournesol et celui d'*Arabidopsis*, qui sont 2 plantes très éloignées mais moins qu'entre *Arabidopsis* et le riz, les remaniements de blocs synténiques soient très nombreux et que la taille de ces blocs soit petite, de l'ordre de quelques Kbases.

La taille du génome du tournesol ne facilite pas les analyses. Les différentes études de synténie faîte sont réalisées entre plantes de petit génome (*Arabidopsis* : 125 Mbases, *Medicago* : $\tilde{5}$00 Mbases, riz : 450 Mbases). Le tournesol, lui, possède un grand génome, 3 000 Mbases, ce qui le rapproche, en taille de génome, du maïs, une plante monocotylédone tout comme le riz. Des études de synténie entre le génome du riz et du maïs ont montré qu'elle était bien conservé, malgré quelque remaniement et la perte ou l'insertion de gènes au sein de bloc syntéique. En revanche, il est flagrant, lorsque l'on compare les régions synténiques du génome du maïs avec celles du riz, que la distance entre les gènes du maïs est beaucoup plus importante que celle observée entre les gènes du riz. Environ 90% du génome du maïs est constitué de séquences répétées (transposons et rétro-transposons). Il est possible que le génome du tournesol contiennent énormément de séquences répétées et que la distance entre les gènes soient beaucoup plus importante que celle observée chez *Arabidopsis*, ce qui ne facilite pas l'étude de la synténie entr eles 2 espèces.

Chapitre 4

Conclusion Générale & Perspectives

La méthodologie mise au point pour analyser le génome du tournesol fonctionne efficacement. Iccare permet de définir des sondes Overgo ou des couples d'amorces efficaces en transférant les informations des gènes d'*Arabidopsis* aux séquences EST ou ARNm et Synteny Search permet de faciliter l'interprétation des résultats expérimentaux.

Les sondes Overgo définies dans des régions exoniques conservées permettent d'identifier les clones BAC qui contiennent la séquence avec un taux de réussite voisin de 70%. Les amorces définies à partir des informations fournies par Iccare permettent d'amplifier les différents cultivars de l'espèces avec un taux de réussite voisin de 90%. La détection du polymorphisme de taille sur agarose n'est que de 30%, mais c'est une technique rapide et facile à mettre en oeuvre. D'autres techniques de détection de polymorphisme pourraient être envisagées : détection par SSCP à condition de ne pas avoir de profil polymorphe aussi complexe que celui de la rubisco (avoir moins de 3 gènes paralogues dans le génome), détection par SNP avec des amorces marquées aux fluorochromes et détection sur séquenceur capillaire, détection par digestion enzymatique des fragments amplifiés (marqueurs CAPS) comme ceux utilisés dans la publication de la carte génétique de F2 (Gentzbittel *et al*, 1999) ou d'autres techniques.

Les différents outils mis en place nous ont permis d'analyser le génome du tournesol et d'étudier la synténie entre le tournesol et *Arabidopsis*. Nous avons pu mettre en place une méthodologie d'analyse du génome du tournesol efficace en permettant la définition de couples d'amorces à partir de séquences EST ou ARNm qui permettent la construction de cartes génétiques et l'analyse et le séquençage des introns des gènes mais aussi la définition de sondes Overgo qui permettent d'identifier des clones BAC contenant les gènes pour la construction de cartes physiques.

Les résultats d'hybridation de la banque de clones BAC nous a permis d'identifier les clones BAC contenant les gènes spécifiques des sondes Overgo mais ces clones n'ont pu être regroupés en contig. Ces résultats ne permettent pas d'exclure ni de confirmer l'hypothèse formulée en début de thèse qui était de savoir si l'organisation du génome du tournesol se rapprocherait plus de celle d'*Arabidopsis* avec des espaces intergéniques plus important ou de

celle du génome du maïs avec des gènes regroupés en îlots. Cependant, le fait que l'on ait pas réussi à regrouper les clones BAC en contig (et si la synténie est conservée entre le génome du tournesol et celui d'*Arabidopsis*) laisse à penser que les espaces intergéniques chez le tournesol sont plus importants que chez *Arabidopsis*, ce qui concorderait avec la première hypothèse. Afin de pouvoir répondre à cette question, de nouveaux résultats doivent être obtenus en analysant quelques clones BAC tout en essayant de les regrouper en comblant les espaces vacants entre eux.

Définir des couples d'amorces de part et d'autres des introns nous a permis d'obtenir de nouveaux marqueurs moléculaires. Ces marqueurs moléculaires ont facilement été cartographiés sur des cartes génétiques. Cependant, le faible nombre de données, autant les nouveaux marqueurs moléculaires définis à partir de Iccare que les séquences EST utilisés pour construire la carte génétique de Gentzbittel *et al.* de 1999, ne nous a pas permis d'établir si la sytnénie était conservée entre le génome du tournesol et celui d'*Arabidopsis*. Il en va de même avec les résultats de l'hybridation de la banque de clones BAC, en revanche ces résultats semblent indiquer que la distance entre les gènes du tournesol est plus importante chez le tournesol que chez *Arabidopsis* et qu'ils ne semblent pas organisés en îlots au sein du génome du tournesol. Ces différents outils et résultats nous ont aussi permis d'estimer le nombre de gènes contenu dans le génome du tournesol à 40 000 / 50 000 gènes.

Tous ces résultats ne sont que préliminaires et bien d'autres peuvent être obtenus. Ce travail de thèse nous a principalement permis de mettre au point la méthode d'analyse du génome du tournesol. Durant la thèse, nous n'avons exploité que 200 gènes supposés orthologues entre le tournesol et *Arabidopsis* (161 pour les sondes Overgo et 51 pour les amorces dont certains communs pour les 2) mais il en reste encore 3 435 qui sont exploitables. Pour la cartographie génétique, en considérant que parmi ces 3 435 séquences, 10% ne permettent pas d'amplification et qu'ensuite entre 15 et 30% donnent un polymorphisme de taille sur agarose, alors nous disposons d'un nombre de marqueurs moléculaires compris entre 468 et 935. Pour le criblage de la banque de clones BAC, nous avons au minimum 2 425 sondes Overgo qui s'hybrideraient avec des clones BAC (en considérant que 70% des sondes soient efficaces). De plus, pour des gènes d'*Arabidopsis* qui ne présenteraient aucune similitude avec des séquences EST ou ARNm du tournesol, nous pourrions tout de même utiliser les séquences d'autres organismes pour définir des sondes Overgo, même si les chances de résultats ne sont que de 1 sur 4.

Ce travail de thèse nous a montré qu'il était difficile d'analyser l'organisation du génome du tournesol malgré des couples d'amorces et des sondes Overgo efficaces. L'analyse de l'organisation du génome du tournesol est rendue difficile par toutes les relations complexes existant entre les gènes. Les duplications, les familles multigéniques, les réarrangements et remaniements chromosomiques ne facilitent pas ce genre d'étude. Il est donc nécessaire de recueillir le maximum d'informations sur les gènes que l'on exploite. Le développement de Synteny Search et l'incorporation des informations des nouveaux génomes séquencés seront une source d'informations essentielle.

L'addition de nouvelles séquences dans les bases de données ainsi que l'arrivée prochaine des séquences complètes de nouveaux génomes tel que le riz, *Medicago truncatula* ou le peuplier devraient permettre d'augmenter le nombre de séquences similaires et d'utiliser ces nouveaux génomes complets comme plantes modèles relais ou nodales dans leurs taxons respectifs afin de compléter les informations fournies par *Arabidopsis*.

Arabidopsis ne possède qu'environ 27 000 gènes, alors que le riz et le peuplier en auraient environ 40 000. Parmi ces gènes surnuméraires vis-à-vis d'*Arabidopsis*, certains sont sûrement des gènes dupliqués de gènes orthologues d'*Arabidopsis*, mais un certain nombre d'entre eux doit être spécifique de la plante. L'important est qu'au fil du temps les informations puissent être cumulées et intégrées afin de pouvoir les exploiter au maximum et de faciliter les analyses de génomes de plantes, comme le tournesol, pour lesquelles les moyens techniques et financiers ne sont pas aussi importants. Déjà, le riz (en cours de finition de séquençage et d'annotation) et *Medicago truncatula* sont utilisés comme plantes nodales pour respectivement les plantes monocotylédones (Goff, 1999) et les plantes légumineuses (Foster-Hartnett *et al.*, 2002, Choi *et al.*, 2004).

D'autres approches pourraient être envisagées pour l'étude de l'organisation du génome du tournseol comme en utilisant des amorces définies à l'aide d'Iccare pour amplifier la banque de clones BAC poolée. En plus des hybridations de la banque de clones BAC avec des sondes Overgo, une stratégie d'amplification de la banque de clones BAC préalablement poolée (pour la stratégie de pool, voir Klein *et al.*, 2000) permettrait d'identifier les clones BAC et d'avoir une confirmation de l'hybridation. Cette stratégie permettrait aussi de rapidement pouvoir combiner les informations de cartographie génétique et de cartographie physique du tournesol.

De plus, afin de faciliter l'analyse de la synténie entre gènes des plantes modèles et du tournesol, il faut utiliser un lot de gènes qui soient bien étudier pour lesquels on connaisse l'ensemble des gènes paralogues, comme pour la rubisco ainsi qu'un lot des gènes que l'on sait être uniques. Ainsi, il sera plus facile de d'analyser la synténie entre gènes de plusieurs organismes végétaux.

En plus des relations complexes qu'il existe entre gènes au sein d'un organisme, la taille du génome du tournesol ne facilite pas l'analyse de son organisation. La stratégie que nous avons utilisée pour analyser l'organisation du génome du tournesol consiste à analyser les régions codantes du génome par l'intermédiaire des séquences EST qui sont le résultat de l'expression des gènes. Mais ces régions codantes ne représentent qu'une faible partie du génome. L'organisation du génome du tournesol est aussi dépendant de ces régions non transcrites. Le tournesol a un génome beaucoup plus grand que celui de la plante modèle *Arabidopsis* ; certes le nombre de gènes du tournesol est estimé entre 40 000 et 50 000 gènes comme pour le riz ou le peuplier, *Arabidopsis* possède environ 27 000 gènes et les gènes surnuméraires par rapport à *Arabidopsis* n'expliquent pas la grande différence de taille de génome chez le tournesol. Cette différence de taille est surtout due aux régions non codantes comme les transposons et rétrotransposons. Une autre voie d'analyse de l'organisation du génome du tournesol serait

donc de s'intéresser à ces séquences. L'analyse des gènes de la rubisco a permis le séquençage d'une séquence atypique pour un gène de la rubisco. Cette séquence ne possède pas d'intron et présente un certain nombre de nucléotides polymorphes (insertion/délétion, substitution) qui lui étaient propres en comparaison des autres séquences des gènes de rubisco. Il est très probable que cette séquence soit le résultat d'une insertion dans le génome d'un ARNm de rubisco par réverse transcription. Nous aurions donc à faire à des rétro-transposons. La piste des transposons et rétro-transposons n'a pas été explorée mais les plantes à grand génome contiennent de grandes quantité d'éléments mobiles et le tournesol ne doit pas échapper à la règle. D'ailleurs, cette hypothèse pourrait très bien aller dans le même sens de l'espace inter-génique bien plus important chez le tournesol que chez *Arabidopsis*.

Chapitre 5

Matériels et Méthodes

5.1 Analyse de la Rubisco

Matériels végétaux

Sept cultivars (ou génotypes) de tournesol (Ha300 noté [A], psc8 [B], KA [C], CP73 [D], Ha821 [E], RHA266 [F] et PAC2 [G]) ont été utilisés pour tester les couples d'amorces et vérifier la présence d'introns dans les gènes *rbcS* du tournesol. Seuls les produits d'amplification F3R1 et F0R4' du cultivar Ha300 ont été séquencés.

Extraction d'ADN

L'extraction d'ADN est réalisée à partir de fragments de feuilles stockées à -20°C d'après la méthode de Fulton *et al.*, 1995 modifiée (voire l'annexe *K* page 187). Les ADN sont repris dans 100 μL de tampon TE et chaque extrait est dosé au spectrophotomètre. La qualité des ADN est vérifiée par électrophorèse sur gel d'agarose 1%, puis chaque extrait est dilué afin d'obtenir une concentration finale de 50 ng.μL^{-1}.

Amplification par PCR

L'amplification des ADN de 7 cultivars de tournesol (plus 1 sans ADN noté H qui servira de témoin négatif) se fait en utilisant la technique de PCR (Polymerase Chain Reaction = Réaction de polymérisation en chaîne). En plus des amorces faites pour la rubisco, des amorces de calmoduline seront utilisées comme témoins positifs d'amplification. Le milieu réactionnel (50 μL final) est composé de 10 mM de Tris-HCl (pH 8.4), de 50 mM de KCl, de 2,5 mM de $MgCl_2$, 2 mM d'une solution des 4 dNTPs, 5 U de Taq polymérase (*Thermus aquaticus*) [GibcoBRL©], 2 pmoles de chaque amorce et 50 ng d'ADN. L'amplification se fait dans un thermocycleur après une étape de dénaturation de 4 min à 94°C, puis un cycle répété 30 fois de 30 sec de dénaturation à 94°C suivi par 30 sec d'hybridation à 56°C et enfin 30 sec d'élongation à 72°C, pour finir par une étape d'élongation de 6 min à 72°C.

Migration sur gel d'agarose

La migration des produits PCR s'effectue par électrophorèse sur des gel d'agarose (2% agarose, 100 mL de TAE 0.5x, BET 50 μg.mL^{-1}) dans du Tris Acétate EDTA 0.5x (TAE) sous l'influence d'un champ électrique de 100 Volts pendant 1 heure. Aux produits d'amplification sont ajoutés 5 μL de tampon de charge. Dix μL de ce mélange sont déposés dans chaque puits du gel. Un μg de marqueur de poids moléculaire est ajouté (100 bp DNA Ladder ou 1 kb DNA Ladder de BioLabs). La visualisation des fragments d'ADN se fait sous UV à l'aide d'un transilluminateur ($\nu = 312$ nm).

Recherche de ploymorphisme par SSCP

L'analyse SSCP va permettre de déterminer le nombre de gènes *rbcS* existant chez différents cultivars de tournesol. Les produits d'amplification de F1R1, F3R1, F3R3 et F3R4 pour tous les cultivars sont passés sur gel de SSCP. Le Stacking gel et le Separating gel sont composés de polyacrylamide à respectivement 6 et 9% (pour plus de détails les gels de polyacrylamide pour SSCP, se reporter à l'annexe *L* page 188). Une fois le gel polymérisé, un pré-run de 15 min aux mêmes conditions de migration (voltage identique) est lancé avec du bleu de migration. Durant le pré-run, 10 μL de produits d'amplification sont repris dans 10 μL de tampon de charge "Bleu-formamide" puis dénaturés à 94°C pendant 5 min et placés sur glace pendant la durée du pré-run. Les échantillons sont déposés sur le gel. La migration est réalisée dans du tampon TBE (Tris Borate EDTA) 0.5x à 4°C (en chambre froide) pendant 16 heures minimum sous un voltage propre à chaque gel utilisé ainsi qu'à la taille des fragments amplifiés. Le voltage est proportionnel à la taille des fragments ; sur petit gel le voltage est de 1 V par paire de base, alors que sur grand gel, il est de 2 V par paire de bases. Pour des fragments de 100 pb, le voltage appliqué sera de 100 V sur petit gel et de 200 V sur grand gel.

La révélation du gel se fait au nitrate d'argent. Le gel est décollé des plaques de verre et mis en présence d'une solution d'éthanol 10% pendant 5 minutes (fixation). Une solution d'acide nitrique 1% permet d'oxyder l'ADN (3 minutes). Le gel est ensuite rincé deux fois à l'eau distillée puis coloré au nitrate d'argent (2 g.L^{-1}) pendant 20 minutes à l'obscurité. Le gel est de nouveau rincé deux fois à l'eau distillée puis révélé par ajout d'une solution de révélation (carbonate de sodium (30 g.L^{-1}) additionnée de 270 μL de formadéhyde pour 500 mL). Laisser agiter jusqu'à la visualisation des bandes sur le gel. Stopper la réaction avec de l'acide acétique 10% pendant 15 minutes. Un bain de séchage (éthanol 10%, acide acétique 7,5% et glycérol 1%) est réalisé pendant 15 minutes puis le gel est déposé sur papier Whatmann 3MM pour être séché sous vide pendant 1 heure à 80°C.

Clonage et séquençage des produits d'amplification

Les produits d'amplification de F3R1 et F0R4' sur l'ADN génomique de Ha300 ont été séquencés afin d'obtenir les différents types de polymorphisme existant pour ce cultivar mais aussi pour obtenir la séquence du premier intron. F0R4' a été préféré à F0R4 car il semble qu'il y ait un polymorphisme en position +180. Même si R4' possède une base variable (*A*, *T*, *C* et *G*) en position +180, les résultats de l'amplification de F3R1 permettront de vérifier qu'il n'y a pas d'incidence particulière avec cette amorce. Les produits d'amplification PCR issus

des amorces F3R1 et F0R4' du cultivar Ha300 sont clonés dans un vecteur pCRII (résistance à l'ampicilline) selon le protocole du kit "TA cloning Invitrogen". Dix μL de produits de ligation sont dessalés par dialyse sur filtre de 0,025 μm (Millipore) puis mélangés à 40 μL de bactéries compétentes (souches DH5α). Le mélange subit une électroporation (impulsion 1,5 kV, capacité de 25 μF et 200 Ω) puis 1 mL de SOC est ajouté au mélange (SOC : 20 g Bacto Tryptone, 5 g Yeast extract, 0,58 g $NaCl$, 0,186 g KCl, complétés à 1 litre avec eau UHQ et autoclaver). Le mélange est mis en culture 1 heure à 37°C sous agitation. Les produits transformés sont ensuite étalés sur milieu LB (ampicilline) additionné de substrat X-gal et mis en culture une nuit à 37°C. Les colonies blanches sont alors repiquées et un "boiling" est réalisé pour vérifier la présence des inserts (même condition d'amplification PCR qu'indiqué page 120 à l'exception des amorces qui sont des amorces M13F et M13R). Les clones positifs sont alors purifiés par miniprep (voir annexe *M* page 190) puis séquencés en utilisant le kit "Prism Amplitaq FS Big Dye Terminators DichloroRhodamine V3" en présence de 5 μL de plasmide pour un volume total de 10 μL. Les clones sont séquencées en 5' et en 3' en utilisant respectivement les amorces M13F et M13R. La purification des produits de séquençage est réalisée grâce au kit Millipore "Montage Seq96" (Sequencing Reaction Clean Up Kit) et l'analyse de la réaction de séquence est faite sur un séquenceur APPLERA 3700. Les données sont ensuite récupérées et traitées afin d'éliminer les séquences du vecteur pour chacune des séquences. Les séquences Forward et Reverse sont ensuite alignées pour vérifier la qualité du séquençage. Au total, 24 et 28 séquences, notées respectivement AIII pour F3R1 et AVII pour F0R4' suivi du numéro de clone, ont été obtenues.

Criblage de la banque de clones BAC

Les différentes sondes Overgo spécifiques du gène *rbcS* sont hybridées sur la banque de clones BAC (pour plus de détails voir page 125).

Extraction et amplification d'ADN de clones BAC

Les ADN des clones BAC positifs aux sondes du gène *rbcS* sont extraits (pour plus de détails voire page 126). Les amorces F0R4 sont utilisées pour vérifier la présence du gène *rbcS* sur les différents clones BAC (pour plus de détails sur les conditions de PCR voir page 126).

5.2 Criblage de la banque de clones BAC

5.2.1 Matériels & Méthodes

Définir des Sondes Overgo

Les sondes Overgo ont été définies à partir de séquences EST ou ARNm de tournesol (et quelque unes à partir de séquences d'autres organismes) qui présentent des similitudes avec des gènes d'*Arabidopsis* localisés sur le chromosome 5. Deux régions particulières ont été étudiées. La première est la région qui contient 3 des 4 gènes *rbcS* positionnés en tandem sur le chromosome 5, la seconde est une région de 7 Mbases sur le haut du chromosome 5.

La région de la rubisco exploitée de 1,34 Mbases est en position $+14,91$ à $+16,25$ Mbases du chromosome 5 et l'autre région exploitée de 7 Mbases est en position 0 à 7 Mbases du chromosome 5.

Dans la première région, 12 sondes ont été définies dont une pour la rubisco et ont été numérotées r1 à r12 (voir Annexe K page 191). Dans la deuxième région, 149 sondes ont été définies et numérotées de 1 à 151, les numéros 29 et 31 n'ont pas été utilisés (voir l'annexe *N* page 191). Trois types différents de sondes Overgo ont été définies. Les premier et deuxième types de sondes correspondent à des sondes Overgo définies à partir de séquences EST ou ARNm de tournesol ayant respectivement de *fortes* et de *faibles similitudes*. Ces sondes Overgo sont appelées sondes Overgo e^{-20} et sondes Overgo e^{-5}. Le dernier type de sondes correspond à des sondes Overgo définies à partir de séquences EST ou ARNm d'organismes autres que le tournesol mais dans des régions ayant de fortes similitudes avec au moins quatre organismes différents dont *Arabidopsis*. Elles sont appelées sondes Overgo *multi* et la séquence utilisée pour définir les sondes Overgo est généralement une séquence EST de *Medicago truncatula*. Les séquences des différents organismes sont comparées entre elles et les nucléotides les plus fréquents sont utilisés pour définir la sonde (pour plus de détails voir l'annexe *O* page 194). La position des *gènes similaires* utilisés est représentée Figure 5.1. Les gènes en bleu ne sont pas utilisés, ceux en rouge présentent de fortes similitudes avec des séquences de tournesol (Overgo e^{-20}), ceux en orange présentent de faibles simlitudes avec des séquences de tournesol (Overgo e^{-5}), et ceux en vert présentent de fortes ou faibles similitudes avec des séquences d'au moins trois organismes différents (Overgo de type *multi*). Au final, nous disposons de 115 sondes Overgo e^{-20}, de 35 sondes Overgo e^{-5} et de 11 sondes Overgo *multi*.

Duplication et repiquage de la banque de clones BAC sur Filtre Haute Densité

La banque de clones BAC utilisée a été construite par Gentzbittel *et al.* en 2002 à partir de la lignée Ha821 par une digestion partielle du génome avec *Hind*III. Avant son utilisation, la banque de clones BAC a d'abord été répliquée à l'aide des outils de la plateforme du Génopôle Midi-Pyrénées de Toulouse dans des plaques Genetix au format 384 contenant du LB freezing additionné de chloramphenicol (12,5 mg.L^{-1}). Des filtres Haute Densité ont ensuite été préparés en utilisant les 408 plaques format 384 de cette banque de clones BAC Ha821/*Hind*III. Un motif de dépôt en 4x4 qui permet le dépôt de 48 plaques 384 en double dépôt sur de simple membrane nylon Immobilon N+ (Millipore) a été utilisé (pour plus de détails voir l'annexe *E* page 164). La banque de clones BAC tient sur 9 membranes (notées de *A* à *I*), qui constitue donc un lot de membranes. Seize lots de membranes ont ainsi été créés. Après le dépôt, les membranes ont prudemment été positionnées sur boîte 22x22 contenant du milieu Luria-Bertani (LB) solide (additionné de 12,5 μg.mL^{-1} de chloramphenicol) avec les bactéries vers le haut. Les boîtes/membranes ont été couvertes, retournées et placées à 37°C pendant 10 à 12 heures. Les membranes ont été retirées et dénaturées (1,5 M $NaCl$ et 0,5 M $NaOH$) pendant 7 minutes, puis neutralisées (1,5 M $NaCl$ et 0,5 M TrisHCl) pendant 7 minutes. Les membranes sont séchées à température ambiante pendant 1 heure. Elles sont ensuite traitées (0,4 M $NaOH$) pendant 20 minutes et rincées (5x SSPE) pendant 7 minutes. Les membranes sont séchées durant la nuit et fixées au Stratalinker. Les membranes sont

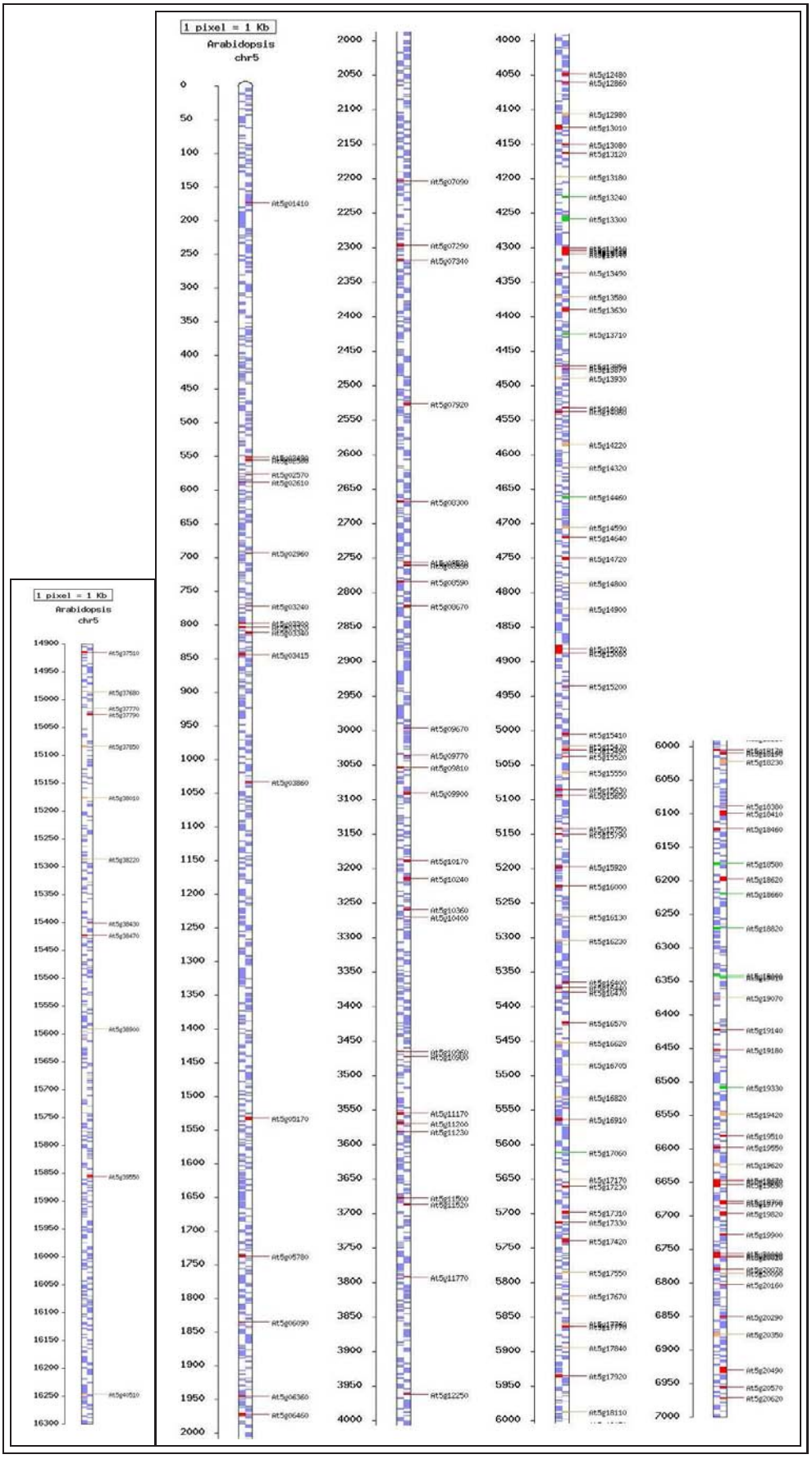


FIG. 5.1 – Position des gènes similaires d'*Arabidopsis* aux séquences EST ou ARNm de tournesol exploitées sur le chromosome 5 d'*Arabidopsis*.

stockées à 4°C.

Les clones BAC positifs après l'hybridation des sondes Overgo en pool sont repiqués et déposés sur de nouvelles membranes 11x7,5 (format d'une plaque 384). En plus des clones BAC positifs, des clones BAC qui n'ont pas été hybridés par les sondes Overgo sont utilisés sur la membrane comme témoins négatifs d'hybridation alors que les clones BAC positifs à la sonde Overgo de rubisco sont utilisés comme clones BAC positifs d'hybridation. Tous les clones BAC sont déposés sur les membranes suivant un pattern qui permet l'identification des clones (voir l'annexe *P* page 196).

Marquage des sondes Overgo

Les sondes Overgo sont dénaturées en combinant 10 pmol de chaque oligo pour un volume final de 2 μL qui a été chauffé dans un thermocycler à 80°C pendant 5 minutes, puis 37°C pendant 10 minutes et enfin refroidi à 4°C. Les sondes Overgos sont marquées individuellement au [α^{-32}P]dCTP et [α^{-32}P]dATP. A chaque sonde dénaturée sont ajouté 8 μL de tampon de marquage pour un marquage [0,5 μL BSA (2mg/mL), 2 μL OLB (1 :2.5 :1.5 ; solution A : 1 mL TrisHCl, pH 8.0 1,25 M, 18 μL 2-mercaptoethanol, 5 μL dGTP 0,1 M, 5 μL dTTP 0,1 M ; solution B : 2 M HEPES-$NaOH$, pH 6.6 ; solution C : 3 mM TrisHCl, pH 7.4, 0,2 mM EDTA), 0,2 μL [α^{-32}P]dCTP, 0,2 μL [α^{-32}P]dATP, 0,2 μL Klenow (5U/μL)] et le tout est incubé à température ambiante pendant 1 heure. Après avoir éliminé les nucléotides non-incorporés en utilisant des colonnes Sephadex G50 (Amersham), les sondes sont regoupées en "pool" et dénaturées à 95°C pendant 10 minutes, puis ajoutées aux tubes d'hybridation contenant les membranes pré-hybridées. Un marquage sert à l'hybridation de deux membranes.

Criblage de la banque de clones BAC

L'hybridation se passe en deux étapes. La première consiste à hybrider sur la totalité de la banque de clones BAC des pools de sondes Overgo (48, 96 ou 149 sondes en même temps) pour récupérer tous les clones BAC positifs. Chaque membrane est imbibée dans du 2x SSC. Deux membranes, séparées par une pièce de mailles de nylon, sont enroulées avec l'ADN tourné vers l'intérieur et placées dans des tubes d'hybridation de verre de 38 mm sur 250 mm contenant 44 mL de solution d'hybridation (1% BSA, 1 mM EDTA, 7% SDS, 0,5 M disodium phosphate). Les membranes sont pré-hybridées à 58°C avec rotation constante pendant 2 heures. Les sondes Overgos sont dénaturées à 90°C pendant 10 minutes. Ensuite, les pools de sondes sont ajoutées à 6 mL de solution d'hybridation qui sont immédiatement ajoutés au 44 mL de solution d'hybridation des tubes pré-hybridés. L'hybridation se fait à 58°C pendant 14 à 18 heures. Les membranes sont rincées progressivement durant 1 heure à 58°C avec respectivement une solution de 2x SSC, 0,1% SDS (rinçage 1), puis 1,5x SSC, 0,1%SDS (rinçage 2) et enfin 0,5x SSC, 0,1% SDS (rinçage3). Après chaque rinçage, le contraste de radiation des membranes est contrôlé au compteur Geiger-Müller. Les rinçages sont arrêtés lorsque le contraste est suffisant. Les membranes sont alors sellées dans des sacs plastiques polyester Kapak et exposées à des films autoradiographique X-ray (Amersham Biosciences Hyperfilm^{TM} MP) à -80°C pendant 5 jours et demi. Chaque clone BAC positif est identifié grâce au pattern de dépôt. Les membranes sont ensuite analysées et tous les clones BAC

positifs sont alors dupliqués dans de nouvelles plaques 384. Ces plaques sont déposées sur de nouvelles membranes selon un pattern spécifique. L'hybridation de ces nouvelles membranes se fait dans les mêmes conditions que la première, sauf que les sondes ne sont en pool mais par 2 : une sonde + la sonde de Rubisco. La première sonde est l'une des sondes du pool précédement utilisé et la deuxième sonde est la sonde Overgo Rub2 spécifique de la Rubisco. La sonde Rub2 permet de distinguer les extrémités de la membrane pour faciliter la lecture (pour plus de détails voir l'annexe *P* page 196). Deux rinçages de 20 minutes avec du 2x SSC, 0,1% SDS et du 1x SSC, 0,1% SDS sont effectués avant de mettre les membranes dans les sacs plastiques polyester Kapak puis exposées à des films autoradiographiques X-ray à -80°C pendant 5 jours et demi. Les clones BAC positifs sont ensuite identifiés en fonction du motif de dépôt utilisé sur les membranes. Ces hybridations servent à identifier les clones BAC spécifiques d'une sonde.

Chacune des sondes est hybridée au moins deux fois en pool sur les membranes de la banque BAC, puis au moins deux fois sur les membranes contenant les clones BAC positifs afin de vérifier les résultats et d'éviter les erreurs d'interprétation.

Extraction d'ADN de clones BAC

L'extraction d'ADN est réalisée à partir d'une culture de clone BAC de 5 mL de LB+chloramphenicol [12,5 mg.L^{-1}] (voir l'annexe *Q* page 198). Les ADN sont repris dans 50 μL de tampon TE et chaque extrait est dosé au spectrophotomètre. Ils sont ensuite dilués afin d'obtenir une concentration finale de 50 ng.μL^{-1}.

Amplification d'ADN de clones BAC

L'amplification des ADN des clones BAC se fait en utilisant la technique de PCR. La réaction de PCR se fait dans un milieu réactionnel de 20 μL composé de 10 mM de Tris-HCl (pH 8.4), de 50 mM de KCl, de 2,5 mM de $MgCl_2$, 2 mM d'une solution des 4 dNTPs, 2 U de Taq polymérase (*Thermus aquaticus*) [GibcoBRL©], 2 pmoles de chaque amorce et 20 ng d'ADN. L'amplification se fait dans un thermocycleur après une étape de dénaturation de 4 min à 94°C, puis un cycle répété 30 fois de 30 sec de dénaturation à 94°C suivi par 30 sec d'hybridation à 60°C et enfin 30 sec d'élongation à 72°C, pour finir par une étape d'élongation de 6 min à 72°C. Seuls 13 couples d'amorces (P05 à P14) issus des séquences EST sont testés (BQ914752, BQ969550, CD850855, BQ915800, BU025288, CD855786, BU027121, CD854117, BQ976133, CD845622, BQ967599, BQ967098, CD854509, voir l'annexe *R* page 199).

Migration des produits d'amplification

La migration des produits PCR s'effectue dans des cuves électrophorèses RAGE (Rapide Agarose Gel Electrophoresis) contenant du TAE 0.5x sur des gels d'agarose (4% agarose, 50 mL de Tris Acétate EDTA 0.5x, BET 50 μg.mL^{-1}) sous l'influence d'un champ électrique de 160 Volts pendant 20 minutes. Aux produits d'amplification sont ajoutés 2 μL de tampon de charge. Dix μL de ce mélange sont déposés dans chaque puits du gel. Un μg de marqueur de poids moléculaire est ajouté (2 log Ladder de BioLabs). La visualisation des fragments d'ADN se fait sous UV à l'aide d'un transilluminateur ($\nu = 312$ nm).

Fingerprint des clones BAC

L'ADN de 38 clones BAC a été digéré et séparé sur gel d'agarose afin d'obtenir leur " empreinte " (ou fingerprint). Les clones BAC spécifiques de la sonde 39, soit 19 clones BAC, sont digérés et leurs empreintes sont déterminées afin de voir si parmi ces clones certains sont chevauchant. Les clones BAC spéqifiques de 8 sondes supposées proches (141, 114, 71, 115, 72, 116, 142 et 143), soit un total de 19 clones BAC, sont digérés et leurs empreintes sont déterminées afin de voir s'il certains clones sont recouvrants. Pour cela, 10 μg d'ADN de clone BAC sont digérés dans un volume final de 10 μL contenant du tampon 1x, 50 U d'enzyme *Hind*III. La digestion est effectuée à 37°C pendant au moins 4 heures, puis les produits de digestion sont chauffés à 72°C pendant 10 minutes pour désactiver l'enzyme. Les produits de digestion sont alors mis à migrer dans un gel d'agarose 1% dans une solution de TAE 0,5x pendant 17 heures à 20 Volt. Le gel est ensuite coloré dans un bain de BET (50 μg.mL^{-1}) pendant 20 minutes. La visualisation des fragments d'ADN se fait sous UV à l'aide d'un transilluminateur (ν = 312 nm).

5.3 Tester les amorces et Définition de Cartes Génétiques

5.3.1 Matériels & Méthodes

Analyse avec Iccare

Les séquences EST qui ont servi à la construction d'une carte génétique publiée par Gentzbittel *et al.* en 1999 sont comparées aux séquences codantes des gènes d'*Arabidopsis* pour déterminer s'il existe de la macrosynténie entre les deux organismes. Cent vingt et une séquences EST sur les 170 séquences qui ont servi à la construction de la carte génétique ont été recupérées et ont été soumises à Iccare afin d'être comparées aux séquences codantes des gènes d'*Arabidopsis*. Ces séquences EST ont servi à cartographier une population F2 issue du croisement de PAC2 par RHA266. Les 121 séquences utilisées correspondent à 177 locis répartis sur 17 groupes de liaisons (certaines séquences EST sont localisées à plusieurs endroits).

Matériels végétaux

Les ADN de 5 cultivars de tournesol sont utilisés pour chercher du polymorphisme de taille par amplification PCR entre ces différents cultivars (cultivars : PAC2, RHA266, psc8, AS613 et SD1'). Les amorces qui sont polymorphes pour les cultivars PAC2 et RHA266, parents des lignées recombinantes (RILs), seront utilisées pour amplifier les ADN de 93 lignées recombinantes afin de construire une carte génétique.

Définition des couples d'amorces

Les couples d'amorces sont définis à partir de séquences EST ou ARNm de tournesol de *fortes similitudes*. Les gènes d'*Arabidopsis* similaires aux séquences utilisées sont localisés sur le chromosome 5 d'*Arabidopsis* et espacés les uns des autres d'environ 500 Kbases. Les amorces

ont été définies afin de pouvoir amplifier les régions introniques des séquences génomiques pour favoriser le polymorphisme de taille. Toutes les amorces ont été définies dans les exons des séquences EST ayant de fortes similitudes avec les séquences codantes des gènes d'*Arabidopsis* (avec une *E*-value inférieure à $1e^{-20}$). Les séquences des amorces sont présentées dans l'annexe *R* page 199. Les positions des gènes d'*Arabidopsis* sont présentées Figure 5.2. Au total, 51 couples d'amorces ont été définis avec un écart moyen entre gènes chez *Arabidopsis* d'environ 500 Kbases (l'un des couples d'amorces est spécifique du gène de la rubisco At5g38430).

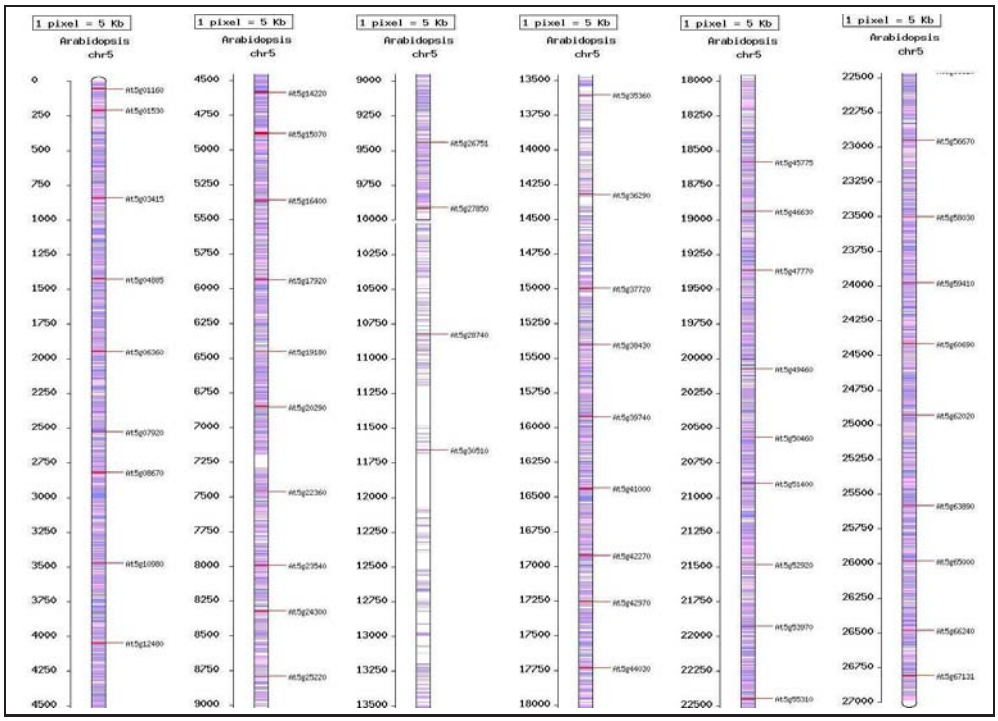


FIG. 5.2 – Position des gènes d'*Arabidopsis* similaires aux séquences de tournesol qui servent à la définition des amorces.

Amplification par PCR

L'amplification des ADN des différents cultivars de tournesol se fait en utilisant la technique de PCR. La réaction de PCR se fait dans un milieu réactionnel final de 50 μL (pour le détail voir page 120).

Migration des produits d'amplification

La migration des produits PCR s'effectue dans des cuves électrophorèses RAGE (pour plus de détails voire page 126). Des gels d'agarose de 2% sont utilisés pour séparer les produits d'amplification des différents cultivars de tournesol. Des gels d'agarose de 4% sont utilisés pour séparer les différents produits d'amplification des lignées recombinantes et la durée de migration passe de 20 minutes à 30 minutes.

Construction de la carte génétique

Le polymorphisme des produits d'amplification des ADN des lignées recombinantes est analysé pour 7 couples d'amorces (P4, P8, P10, P15, P25, P40 et P45). Un fichier contenant

les données de polymorphisme est formaté pour être analysé par MapMaker/Exp V3.0 (Lander *et al.*, 1987). Les différents marqueurs sont analysés avec un LOD score de 3 ou 2 et un pourcentage de recombinaison de 50%. Les marqueurs moléculaires sont ensuite ajoutés aux fichiers de données contenant les marqueurs moléculaires qui ont permis de construire une carte génétique publiée par Rachid Al-chaarani *et al.* en 2004. Cette carte génétique est issue de l'analyse de lignées recombinantes issues du croisement PAC2xRHA266. La carte génétique est constituée de 367 marqueurs auxquels ont été ajoutés 44 nouveaux marqueurs ainsi que nos 7 marqueurs. Les liaisons entre marqueurs sont déterminées par MapMaker avec une valeur de LOD équivalente à celle qui a servi à la construction de la carte publiée (LOD de 4.0).

Bibliographie

A

Acarkan, A., Roßberg, M., Koch, M., Schmidt, R. (2000) Comparative genome analysis reveals extensive conservation of genome organization for *Arabidopsis thaliana* and *Capsella rubella*. *The Plant Journal*, **23(1)**, **55-62**.

Adams, M.D., Celniker, S.E., Holt, Evans, C.A., R.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., *et al.* (2000) The Genome Sequence of Drosophila melanogaster. *Science*, **287**, **2185-2195**.

Adams, K.L., Cronn, R., Percifield, R., Wendel, J.F. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci U S A*, **100(8)**, **4649-4654**.

Adams, K.L., Percifield, R., Wendel, J.F. (2004) Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* **168(4)**, **2217-2226**.

Adams, K.L., Wendel, J.F. (2005) Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* **8**, **135-141**.

Ahn, S., Anderson, JA., Sorrells, M.E., Tanksley, S.D. (1993) Homoeologous relationships of rice, wheat and maize chromosomes. *Mol Gen Genet.*, **241(5-6)**, **483-490**.

Altschul, S.F., Madden, T.L., Schaeffer, A.A., Zhang, J., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Res.*, **25(17)**, **3389-3402**.

Arabidopsis Genome Initiative, The (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, **796-815**.

Arnason, U., Xu, X., Gullberg, A., Graur, D. (1996) The " Phoca Standard " : An External Molecular Reference for Calibrating Recent Evolutionary Divergences. *J. Mol. Evol.*, **43**, **41-45**.

Arnold, M.L. (2004) Transfer and Origin of Adaptations through Natural Hybridization : Were Ander-

son and Stebbins Right ? *The Plant Cell*, **16**, **562-570**.

B

Babula, D., Kaczmarek, M., Barakat, A., Delseny, M., Quiros, C.F., Sadowski, J. (2003) Chromosomal mapping of *Brassica oleracea* based on ESTs from *Arabidopsis thaliana* : complexity of the comparative map. *Mol Gen Genomics*, **268**, **656-665**.

Bancroft, I. (2000) Insights into the structural and functional evolution of plant genomes afforded by the nucleotide sequences of chromosomes 2 and 4 of *Arabidopsis thaliana*. *Yeast*, **17**, **1-5**.

Bancroft, I. (2001) Duplicate and diverge : the evolution of plant genome microstructure. *Trends in Genetics*, **17(2)**, **89-93**.

Barnes, S. (2002) Comparing *Arabidopsis* to other flowering plant. *Current Opinion in Plant Biology*, **5**, **128-133**.

Bennetzen, J.L., Freeling, M. (1997) The unified grass genome : synergy in synteny. *Genome Res.*, **7(4)**, **301-206**.

Bennetzen, J.L., SanMiguel, P., Chen, M., Tikhonov, A., Francki, M., Avramova, Z. (1998) Grass genomes. *PNAS*, **95**, **1975-1978**.

Bennetzen, J.L, Ramakrishna, W. (2002) Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Mol Biol*, **48**, **821-827**.

Bennetzen, J.L., Coleman, C., Liu, R., Ma, J., Ramakrishna, W. (2004) COnsistent over-estimation of gene number in complex plant genomes. *Current Opinion in Plant Biology*, **7**, **732-736**.

Bennetzen, J.L., Ma, J., Devos, K.M. (2005) Mechamisms of Recent Genome Size Variation in Flowering Plants. *Annals of Botany*, **95**, **127-132**.

Bert, P.F., Jouan, I., Tourvieille de Labrouhe, D., Serre, F., Nicolas, P., Vear, F. (2002) Comparative genetic analysis of quantitative traits in sunflower (*Helianthus annuus* L.) 1. QTL involved in resistance to *Sclerotinia sclerotiorum* and *Diaporthe helianthi*. *Theor Appl Genet*, **105(6-7)**, **985-993**.

Bert, P.-F., Jouan, I., Tourvieille de Labrouhe, D., Serre, F., Philippon, J., Nicolas, P., Vear, F. (2003) Comparative genetic analysis of quantitative traits in sunflower (*Helianthus annuus* L.) 2. Characterisation of QTL involved in developmental and agronomic traits. *Theor Appl Genet*, **107**, **181-189**.

Bert, P.-F., Dechamp-Guillaume, G., Serre, F., Jouan, I., Tourvieille de Labrouhe, D., Nicolas, P., Vear, F. (2004) Comparative genetic analysis of quantitative traits in sunflower (*Helianthus annuus* L.) 3. Characterisation of QTL involved in resistance to *Sclerotinia sclerotiorum* and *Phoma macdonaldii*. *Theor*

Appl Genet, **109**, **865-874**.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., Delseny, M. (2000) Extensive Duplication and Reshuffling in the Arabidopsis Genome. *The Plant Cell*, **12**, **1093-1101**.

Blanc, G., Hokamp, K., Wolfe, K.H. (2003) A Recent Polyploidy Superimposed on Older Large-Scale Duplications in the *Arabidopsis* Genome. *Genome Res.*, **13**, **137-144**.

Blanc, G., Wolfe, K.H. (2004) Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution. *The Plant Cell*, **16**, **1679-1691**.

Boivin, K., Acarkan, A., Mbulu, R.S., Clarenz, O., Schmidt, R. (2004) The Arabidopsis genome sequence as a tool for genome analysis in Brassicaceae. A comparison of the Arabidopsis and Capsella rubella genomes. *Plant Physiol.*, **135(2)**, **735-744**.

Bonierbale, M.W., Pleristed, R.L., Tanksley, S.D. (1988) RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. *Genetics*, **120**, **1095-1103**.

Bouzidi, M.F., Badaoui, S., Cambon, F., Vear, F., Tourvieille de Labrouhe, D., Nicolas, P., Mouzeyar, S. (2002) Molecular analysis of a major locus for resistance to downy mildew in sunflower with specific PCR-based markers. *Theor Appl Genet*, **104(4)**, **592-600**.

Bowers, J.E., Abbey, C., Anderson, S., Chang, C., Draye, X., emphet al. (2003) A High-Density Genetic Recombination Map of Sequence-Tagged Sites for *Sorghum*, as a Framework for Comparative Structural and Evolutionary Genomics of Tropical Grains and Grasses. *Genetics*, **165**, **367-386**.

Brandes, A., Heslop-Harrison, J.S., Kamm, A., Kubis, S., Doudrick, R.L., Schmidt, T. (1997) Comparative analysis of the chromosomal and genomic organization of *Ty*1-*copia*-like retrotransposons in pteridophytes, gymnosperms and angiosperms. *Plant Mol Biol*, **33**, **11-21**.

Burke, J.M., Lai, Z., Salmaso, M., Nakazato, T., Tang, S., Heesacker, A., Knapp, S.J., Rieseberg, L.H. (2004) Comparative Mapping and Rapid Karyotypic Evolution in the Genus Helianthus. *Genetics*, **167**, **449-457**.

C

***C. elegans* Sequencing Consortium (The)** (1998) Genome Sequence of the Nematode *C. elegans* : A Platform for Investigating Biology. *Science*, **282**, **2012-2018**.

Cai, W.-W., Reneker, J., Chow, C.-W., Vaishnav, M., Bradley, A. (1998) An Anchored Framework BAC Map of Mouse Chromosome 11 Assembled Using Multiplex Oligonucleotide Hybridization. *Genomics*, **54**, **387-397**.

Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D., May, G. (2004) The roles of segmental

and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biology*, **4**, **10**.

Carels, N., Bernardi, G. (2000) Two classes of genes in plants. *Genetics*, **154(4)**, **1819-1825**.

Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B., Dubois, I., Dossat, C., Sourdille, P., Joudrier, P., Gautier, M.F., Cattolico, L., Beckert, M., Aubourg, S., Weissenbach, J., Caboche, M., Bernard, M., Leroy, P., Chalhoub, B. (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (tritium and aegilops). *Plant Cell*, **7(4)**, **1033-1045**.

Chao, S., Sharp, P.J., Worland, A.J., Warham, E.J., Koebner, R.M.D., Gale, M.D. (1989) RFLP based genetic maps of wheat homoeologous group 7 chromosomes. *Theor. Appl. Genet.*, **78**, **495-504**.

Cheng, S.V., Nadeau, J.H., Tanzi, R.E., Watkins, P.C., Jagadesh, J., Taylor, B.A., Haines, J.L., Sacchi, N., Gusella, J.F. (1988) Comparative mapping of DNA markers from the familial Alzheimer disease and Down syndrome regions of human chromosome 21 to mouse chromosomes 16 and 17. *PNAS*, **85(16)**, **6032-6036**.

Choi, H.-K., Mun, J.-H., Kim, D.-J., Zhu, H., Baek, J.-M., Mudge, J., Roe, B., Ellis, N., Doyle, J., Kiss, G.B., Young, N.D., Cook, D.R. (2004) Estimating genome conservation between crop and model legume species. *PNAS*, **101(43)**, **15289-15294**.

Chowdhary, B.P., Raudsepp, T., Frönicke, L., Scherthan, H. (1998) Emerging Patterns of Comparative Genome Organization in Some Mammalian Species as Revealed by Zoo-FISH. *Genome Res.*, **8**, **577-589**.

Chromosome 21 mapping and sequencing consortium (The) (2000) The DNA sequence of human chromosome 21. *Nature*, **405**, **311-319**.

Comai, L., Tyagi, A.P., Winter, K., Holmes-Davis, R., Reynolds, S.H., Stevens, Y., Byers, B. (2000) Phenotypic Instability and Rapid Gene Silencing in Newly Formed Arabidopsis Allotetraploids. *The Plant Cell*, **12**, **1551-1567**.

Conley, E.J., Nduati, V., Gonzalez-Hernandez, J.L., Mesfin, A., Trudeau-Spanjers, M., Chao, S., Lazo, G.R., Hummel, D.D., Anderson, O.D., Qi, L.L., Gill, B.S., Echalier, B., Linkiewicz, A.M., Dubcovsky, J., Akhunov, E.D., Dvorak, J., Peng, J.H., Lapitan, N.L., Pathan, M.S., Nguyen, H.T., Ma, X.F., Miftahudin, Gustafson, J.P., Greene, R.A., Sorrells, M.E., Hossain, K.G., Kalavacharla, V., Kianian, S.F., Sidhu, D., Dilbirligi, M., Gill, K.S., Choi, D.W., Fenton, R.D., Close, T.J., McGuire, P.E., Qualset, C.O., Anderson, J.A. (2004) A 2600-locus chromosome bin map of wheat homoeologous group 2 reveals interstitial gene-rich islands and colinearity with rice. *Genetics.*, **168(2)**, **625-637**.

Corpet, F. (1988) Multiple sequence aligment with hierarchical clustering. *Nucleic Acids Res.*, **16(22)**, **10881-10890**.

Craig Venter, J., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.* (2001) The Sequence of the Human Genome. *Science*, **291**, **1304-1351**.

Cronn, M.C., Small, R.L., Wendel, J.F. (1999) Duplicated genes evolve independently efter polyploid formation in cotton. *PNAS*, **96(25)**, **14406-14411**.

D

Dean, C., Elzen van den, P., Tamaki, S., Dunsmuir, P., Bedbrook, J. (1985) Linkage and homology analysis divides the eight genes for the small subunit of petunia ribulose 1,5-bis-phosphate carboxylase into three gene families. *PNAS*, **82**, **4964-4968**.

DeRocher, E.J., Quingley, F., Mache, R., Bohnert, H.J. (1993) The six genes of the Rubisco small subunit multigene family from *Mesembryanthemum crystallinum*, a facultative CAM plant. *Mol Gen Genet*, **239**, **450-462**.

Devos, K.M., Atkinson, M.D., Chinoy, C.N., Liu, C.J., Gale, M.D. (1993) RFLP-based genetic map of the homoeologous group 2 chromosomes of wheat, rye, and barley. *Theor. Appl. Genet.*, **85**, **784-792**.

Devos, K.M., Gale, M.D. (2000) Genome relationships : the grass model in current research. *Plant Cell*, **12**, **637-646**.

Devos, K.M., Brown, J.K.M., Bennetzen, J.L. (2002) Genome Size Reduction through Illegitimate Recombination Counteracts genome Expansion in *Arabidopsis*. *Genome Res.*, **12**, **1075-1079**.

Devos, K.M. (2005) Updating the 'Crop Circle'. *Current Opinion in Plant Biology*, **8**, **155-162**.

Doganlar, S., Frary, A., Daunay, M.C., Lester, R.N., Tanksley, S.D. (2002) A comparative genetic linkage map of eggplant (*Solanum melongena*) and its implications for genome evolution in the solanaceae. *Genetics*, **161**, **1697-1711**.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, **489-496**.

E

Echard, G., Broad, T.E., Hill, D., Pearce, P. (1994) Present status of ovine gene map (Ovis aries) ; comparison with the bovine map (Bos Taurus). *Mamm. Genome*, **5(6)**, **324-332**.

Echenique, V., Stamova, B., Wolters, P., Lazo, G., Carollo, V.L., Dubcovsky, J. (2002) Frequencies of Ty1-*copia* and Ty3-*gypsy* retroelements within the *Triticeae* EST databases. *Theor Appl Genet*, **104**, **840-844**.

Elgar, G., Sandford, R., Aparicio, S., Macrae, A., Venkatesh, B., Brenner, S. (1996) Small is beautiful : comparative genomics with the pufferfish (*Fugu rubripes*). *Trends in Genetics*, **12(4)**, **145-150**.

Ermolaeva, M.D., Wu, M., Eisen, J.A., Salzberg, S.L. (2003) The age of the *Arabidopsis thaliana* genome duplication. *Plant Mol Biol*, **51**, **859-866**.

Esquivel, M.G., Anwaruzzaman, M., Spreitzer, R.J. (2002) Deletion of nine carboxy-terminal residues of the Rubisco small subunit decreases thermal stability but does not eliminate function. *FEBS lett*, **520**, **73-76**.

F

Feng, Q., Zhang, Y., Hao, P., Wang, S., Fu, G., Huang, Y., Li, Y., Zhu, J., Liu, Y., *et al.* (2002) Sequence and analysis of rice chromosome 4. *Nature*, **420(6913)**, **316-320**.

Feuillet, C., Keller, B. (1999) High gene density is conserved at syntenic loci of small and large grass genomes. *PNAS*, **96**, **8265-8270**.

Feuillet, C., Penger, A., Gellner, K., Mast, A., Keller, B. (2001) Molecular evolution of receptor-like kinase genes in hexaploid wheat : independent evolution of orthologs after polyploidization and mechanisms of local rearrangements at paralogous loci. *Plant Physiology*, **125**, **1304-1313**.

Feuillet, C., Keller, B. (2002) Comparatuve Genomics in the Grass Family : Molecular Characterization of Grass Genome Struture ans Evolution. *Annals of Botany*, **89**, **3-10**.

Fitch, W.M. (2000) Homology a personal view on some of the problems. *Trends in Genetics*, **16(5)**, **227-231**.

Flor, H.H. (1971) Current status of the gene for gene concept. *An. Rev. Phytopathol.*, **9**, **275-286**.

Flores-Berrios, E., Sarrafi,A., Fabre, F., Alibert, G., Gentzbittel, L. (2000) Genotypic variation and chromosomal location of QTLs for somatic embryogenesis revealed by epidermal layers culture of recombinant inbred lines in the sunflower (*Helianthus annuus* L.). *Theor Appl Genet*, **101**, **1307-1312**.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, **1531-1545**.

Foster-Hartnett, D., Mudge, J., Larsen, D., Danesh, D., Yan, H., Denny, R., Penuela, S., Young, N.D. (2002) Comparative genomic analysis of sequences sampled from a small region on soybean (*Glycine max*) molecular linkage group G. *Genome*, **45**, **634-645**.

Fourman, M., Barret, P., Froger, N., Baron, C., Charlot, F., Delourme, R., Brunel, D. (2002) From *Arabidopsis thaliana* to *Brassica napus* : development of amplified consensus genetic markers (ACGM) for construction of a gene map. *Theor Appl Genet*, **105**, **1196-1206**.

Fritz, C.C., Wolter, F.P., Schenkemeyer, V., Herget, T., Schreier, P.H. (1993) The gene family encoding the ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) small subunit of potato. *Gene*, **137**,

271-274.

G

Gale, M.D., Devos, K.M. (1998) Comparative genetics in the grasses. *PNAS*, **95**, **1971-1974**.

Gardiner, J., Schroeder, S., Polacco, M.L., Sanchez-Villeda, H., Fang, Z., Morgante, M., *et al.* (2004) Anchoring 9,371 Maize Expressed Sequence Tagged Unigenes to the Bacterial Artificial Chromosome COntig Map by Two-Dimensional Overgo Hybridization. *Plant Physiol*, **134**, **1-10**.

Gaut, B.S., Doebley, J.F. (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. *PNAS*, **94**, **6809-6814**.

Gaut, B.S., Le Thierry d'Ennequin, M., Peek, A.S., Sawkins, M.C. (2000) Maize as a model for the evolution of plant nuclear genomes. *PNAS*, **97(13)**, **7008-7015**.

Gautier, M., Hayes, H., Eggen, A. (2002) An extensive and comprehensive radiation hybrid map of bovine Chromosome 15 : comparison with human Chromosome 11. *Mamm. Genome*, **13(6)**, **316-319**.

Ge, S., Sang, T., Lu, B.-R., Hong, D.-Y. (1999) Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *PNAS*, **96**, **14400-14405**.

Gebhardt, C., Ritter, E., Barone, A., Debener, T., Walkemeier, B., Schachtschnabel, U., Kaufmann, H., Thompson, R.D., Bonierbale, M.W., Ganal, M.W., *et al.* (1991) RFLP maps of potato and their alignement with the homoleogous tomato genome. *Theor Appl Genet*, **83**, **49-57**.

Gebhardt, C., Walkemeier, B., Henselewski, H., Barakat, M., Delseny, M., Stüber, K. (2003) Comparative mapping between potato (*Solanum tuberosum* and *Arabidopsis thaliana* reveals structurally conserved domains and ancient suplications in the potato genome. *The Plant Journal*, **34**, **529-541**.

Gedil, M.A., Slabaugh, M.B., Berry, S., Johnson, R., Michelmore, R., Miller, J., Gulya, T., Knapp, S.J. (2001) Candidate disease resistance genes in sunflower clones using conserved nucleotide-binding site motifs : Genetic mapping and linkage to the downy mildew resistance gene *Pl1*. *Genome*, **44**, **205-212**.

Gentzbittel, L., Vear, F., Zhang, Y.X., Bervillé, A., Nicolas, P. (1995) Development of a consensus linkage RFLP map of cultivated sunflower (*Helianthus annuus* L.). *Theor Appl Genet*, **90**, **1079-1086**.

Gentzbittel, L., Mouzeyar, S., Badaoui, S., Mestries, E., Vear, F., Tourvieille de Labrouhe, D., Nicolas, P. (1998) Cloning of molecular markers for disease resistance in sunflower, *Helianthus annuus* L. *Theor Appl Genet*, **96**, **519-525**.

Gentzbittel, L., Mestries, E., Mouzeyar, S., Mazeyrat, F., Badaoui, S., Vear, F., Tourvieille de Labrouhe, D., Nicolas, P. (1999) A composite map of expressed sequences and phenotypic traits of the sunflower (*Helianthus annuus* L.) genome. *Theor Appl Genet*, **99**, **218-234**.

Gentzbittel, L., Abbott, A., Galaud, J.P., Georgi, L., Fabre, F., Liboz, T., Alibert, G. (2002) A bacterail artificial chromosome (BAC) library for sunflower, and identification of clones containing genes for putative transmembrane receptors. *Mol Gen Genomics*, **266**, **979-987**.

Goff, S.A. (1999) Rice as a model for cereal genomics. *Current Opinion in Plant Biology*, **2**, **86-89**.

Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., *et al.* (2002) A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, **92-100**.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., *et al.* (1996) Life with 6000 Genes. *Science*, **274**, **563-567**.

Goldschmidt-Clermont, M., Rahire, M. (1986) Sequence, evolution and differential expression of the two genes encoding variant small subunits of ribulose bisphosphate carboxylase/oxygenase in *Chlamydomonas reinhardtii*. *J Mol Biol.*, **191(3)**, **421-432**.

Goureau, A., Garrigues, A., Tosser-Klopp, G., Lahbib-Mansais, Y., Chardon, P., Yerle, M. (2001) Conserved synteny and gene order difference between human chromosome 12 and pig chromosome 5. *Cytogenet Cell Genet*, **94(1-2)**, **49-54**.

Grant, V. (1981) *Plant Speciation (Columbia Univ. Press, New York).*

Grant, D., Cregan, P., Schoemaker, R.C. (2000) Genome organization in dicots : Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *PNAS*, **97(8)**, **4168-4173**.

Guignard, J.-L. (2001) Botanique, Systématique moléculaire, 12^{e}éditions. *Editions Masson S.A.*, Paris cedex 6.

Guimarães, C.T., Sills, G.R., Sobral, B.W.S. (1997) Comparative mapping of Andropogoneae : *Saccharum* L. (sugarcane) and its relation to sorghum and maize. *Proc. Natl. Acad. Sci. USA*, **94**, **14261-14266**.

Guyot, R., Keller, B. (2004) Ancestral genome duplication in rice. *Genome*, **47**, **610-614**.

H

Han, C.S., Sutherland, R.D., Jewett, P.B., Campbell, M.L., Meincke, L.J., *et al.* (2000) Construction of a BAC Contig Map of Chromosome 16q by Two-Dimensional Overgo Hybridization. *Genome Res.*, **10**, **714-721**.

Heiser, C.B., Smith, J.D.M. (1965) New chromosome numbers in *Helianthus* and related genera (Compositae). *Proc Ind Acad Sci*, **64**, **250-253**.

Hervé, D., Fabre, F., Flores-Berrios, E., Leroux, N., Rachid Al-Chaarani, G., Planchon, C., Sarrafi, A., Gentzbittel, L. (2001) QTL analysis of photosynthesis and water status traits in sunflower (*Helianthus annuus* L.) under greenhouse conditions. *Journal of Experimental Botany*, **52(362)**, **1-8**.

Heslop-Harrison, J.S., Murata, M., Ogura, Y., Schwarzacher, T., Motoyoshi, F. (1999) Polymorphisms and Genomic Organization of Repetitive DNA from Centromeric Regions of Arabidopsis Chromosomes. *The Plant Cell*, **11**, **31-42**.

Hongtrakul, V., Slabaugh, M.B., Knapp, S.J. (1998) DFLP, SSCP, and SSR markers for δ9-stearoyl-acyl carrier protein desaturases strongly expressed in developing seeds of sunflower : intron lengths are polymorphic among elite inbred lines. *Molecular Breeding*, **4**, **195-203**.

Hulbert, S.H., Richter, T.E., Axtell, J.D., Bennetzen, J.L. (1990) Genetic mapping and characterization of sorghum and related crops by means of maize DNA probes. *Proc. Natl. Acad. Sci. USA*, **87**, **4251-4255**.

I

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, **860-921**.

J

Jan, C.C., Vick, B.A., Miller, J.F., Kahler, A.L., Butler, III, E.T. (1998) Construction of an RFLP linkage map for cultivated sunflower. *Theor Appl Genet*, **96**, **15-22**.

Janke, A., Xu, X., Arnason, U. (1997) The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupiala, and Eutheria. *PNAS*, **94**, **1276-1281**.

Johnston, J.S., Pepper, A.E., Hall, A.E., Chen, Z.J., Hodnett, G., Drabek, J., Lopez, R., Price, H.J. (2005) Evolution of Genome Size in Brassicaceae. *Annals of Botany*, **95**, **229-235**.

K

Kilian, A., Kudrna, D.A., Kleinhofs, A., Yano, M., Kurata, N., Steffenson, B., Sasaki, T. (1995) Rice-barley synteny and its application to saturation mapping of the barley Rpg1 region. *Nucleic Acids Res.*, **23(14)**, **2729-2733**.

Kim, S.-C., Rieseberg, L.H. (1999) Genetic Architecture of Species Differences in Annual Sunflowers : Implications for Adaptive Trait Introgression. *Genetics*, **153**, **965-977**.

Kim, J., Gordon, L., Dehal, P., Badri, H., Christensen, M., Groza, M., Ha, C., Hammond, S., Vargas, M., Wehri, E., Wagner, M., Olsen, A., Stubbs, L. (2001) Homology-Driven Assembly of a Sequence-Ready Mouse BAC Contig Map Spanning Regions Related to the 46-Mb Gene-Rich Euchromatic Segments of Human Chromosome 19. *Genomics*, **74**, **129-141**.

Klein, P.E., Klein, R.R., Cartinhour, S.W., Ulanch, P.E., Dong, J., Obert, J.A., Morishige, D.T., Schlueter, S.D., Childs, K.L., Ale, M., Mullet, J.E. (2000) A High-throughput AFLP-based Method for Constructing Integrated Genetic and Physical Maps ; Progress Toward a Sorghum Genome Map. *genome Res.*, **10**, **789-807**.

Knight, M.R., Jenkins, G.I. (1992) Genes encoding the small subunit of ribulose 1,5-bisphosphate carboxylase/oxygenase in *Phaseolus vulgaris* L. : nucleotide sequence of cDNA clones and initial studies of expression. *Plant Mol Biol.*, **18(3)**, **567-579**.

Kowalski, S.P., Lan, T.-H., Feldmann, K.A., Paterson, A.H. (1994) Comparative Mapping of *Arabidopsis thaliana* and *Brassica oleracea* Chromosomes Reveals Islands of Conserved Organization. *Genetics*, **138**, **499-510**.

Ku, H.-M., Vision, T., Liu, J., Tanksley, S.D. (2000) Comparing sequenced segments of the tomato and *Arabidopsis* genomes : Large-scale duplication followed by selective gene loss creates a network of synteny. *PNAS*, **97(16)**, **9121-9126**.

Ku, H.-M., Liu, J., Doganlar, S., Tanksley, S.D. (2001) Exploitation of *Arabidopsis*-tomato synteny to construct a high-resolution map of the *ovate*-containing region in tomato chromosome 2. *Genome*, **44**, **470-475**.

Kumar, A., Bennetzen, J.L. (1999) Plant retrotransposons. *Annu Rev Genet*, **33**, **479-532**.

Kumekawa, N., Ohtsubo, E., Ohtsubo, H. (1999) Identification and phylogenetic analysis of *gypsy*-type retrotransposons in the plant kingdom. *Genes Genet Syst*, **74**, **299-307**.

Kurata, N., Moore, G., Nagamura, Y., Foote, T., Yano, M., Minobe, Y., Gale, M. (1994) Conservation of genome structure between rice and wheat. *Bio/Technology*, **12**, **276-278**.

L

Lagercrantz, U., Lydiate, D.J. (1996) Comparative genome mapping in Brassica. *Genetics*, **144**, **1903-1910**.

Lagercrantz, U. (1998) Comparative Mapping Between *Arabidopsis thaliana* and *Brassica nigra* Indicates That Brassica Genomes Have Evolved Through Extensive Genome Replication Accompanied by Chromosome Fusions and Frequent Rearrangements. *Genetics*, **150**, **1217-1228**.

Lai, J., Ma, J., Swigonova, Z., Ramakrishna, W., Linton, E., Llaca, V., Tanyolac, B., Park, Y.-J., Jeong, O.-Y., Bennetzen, J.L., Messing, J. (2004) Gene Loss and Movement in the Maize Genome. *Genome Res.*, **14**, **1924-1931**.

Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E., Newburg, L. (1987) MAPMAKER : an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, **1(2)**, **174-181**.

Langar, K., Lorieux, M., Desmarais, E., Griveau, Y., Gentzbittel, L., Bervillé, A. (2003) Combined mapping of DALP and AFLP markers in cultivated sunflower using F9 recombinant inbred lines. *Theor Appl Genet*, **106**, **1068-1074**.

Leclercq, P. (1969) La stérilité mâle cytoplasmique du tournesol. Premières études sur la restauration de la fertilité.. *Ann. Amelio.*, **19**, **99-109**.

Lecointre, G., Le Guyaver, H. (2001) Classification phylogénétique du vivant, 2eédition. *Editions Belin*, Paris Cedex 6.

Leitch, I.J., Bennett, M.D. (1997) Polyploidy in angiosperms. *Trends in plant science*, **2(12)**, **470-476**.

Li, G., Gao, M., Yang, B., Quiros, C.F. (2003) Gene for gene alignment between the Brassica and Arabidopsis genomes by direct transcriptome mapping. *Theor Appl Genet*, **107**, **168-180**.

Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.-I., *et al.* (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*, **402**, **761-772**.

Livingstone, K.D., Lackney, V.K., Blauth, J.R., van Wijk, R., Jahn, M.K. (1999) Genome Mapping in Capsicum and the Evolution of Genome Structure in the Solanaceae. *Genetics*, **152**, **1183-1202**.

Lopez-Corrales, N.L., Sonstegard, T.S., Smith, T.P. (1998) Comparative gene mapping : cytogenetic localization of PROC, EN1, ALPI, TNP1, and IL1B in cattle and sheep reveals a conserved rearrangement relative to the human genome. *Cytogenet Cell Genet*, **83(1-2)**, **35-38**.

Lukens, L., Zou, F., Lydiate, D., Parkin, I., Osborn, T. (2003) Comparison of a *Brassica oleracea* Genetic Map With the Genome of *Arabidopsis thaliana*. *Genetics*, **164**, **359-372**.

Lunn, J.E. (2003) Sucrose-Phosphatase gene families in plants. *Gene*, **303**, **187-196**.

M

Ma, J., Devos, K.M., Bennetzen, J.L. (2004) Analysis of LTR-Retrotransposons Structures Reveal Recent and Rapid Genomic DNA Loss in Rice. *Genome Res.*, **14**, **860-869**.

Martin-Wess, F., Voß-Nemitz, R., Drögemüller, C., Brenig, B., Leeb, T. (2002) Construction of a 1.2-Mb BAC/PAC Contig of the Porcine Gene *RYR1* Region on SSC 6q1.2 and Comparative Analysis with HSA 19Q13.13. *Genomics*, **80**, **416-422**.

Masterson, J. (1994) Stomatal size in fossil plants : evidence for polyploidy in majority of angiosperms. *Science* **264**, **421-424**.

Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., *et al.* (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, **402**, **769-780**.

Mayer, K., Murphy, G., Tarchini, R., Wambutt, R., Volckaert, G., *et al.* (2001) Conservation of Microstructure between a Sequenced Region of the Genome of Rice and Multiple Segments of the Genome of *Arabidopsis thaliana*. *Genome res.*, **11**, **1167-1174**.

Mazur, B.J., Chui, C.-F. (1985) Sequence of a genomic DNA clone for the small subunit of ribulose

bis-phosphate carboxylase-oxygenase from tobacco. *Nucleic Acids Res.*, **13(7)**, **2373-2385**.

McLysaght, A., Enright, A.J., Skrabanek, L., Wolfe, K.H. (2000) Estimation of synteny conservation and genome compaction between puferfish (Fugu) and human. *Yeast*, **17(1)**, **22-36**.

Meyers, B.C., Dickerman, A.W., Michelmore, R.W., Sivaramakrishnan, S., Sobral, B.W., Young, N.D. (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *The Plant Journal*, **20(3)**, **317-332**.

Micic, Z., Hahn, V., Bauer, E., Schon, C.C., Knapp, S.J., Tang, S., Melchinger, A.E. (2004) QTL mapping of Sclerotinia midstalk-rot resistance in sunflower. *Theor Appl Genet*, **109(7)**, **1474-1484**.

Micic, Z., Hahn, V., Bauer, E., Melchinger, A.E., Knapp, S.J., Tang, S., Schon, C.C. (2005) Identification and validation of QTL for Sclerotinia midstalk rot resistance in sunflower by selective genotyping. *Theor Appl Genet*, **Epub ahead of print**.

Micic, Z., Hahn, V., Bauer, E., Schon, C.C., Melchinger, A.E. (2005) QTL mapping of resistance to Sclerotinia midstalk rot in RIL of sunflower population NDBLOSsel x CM625. *Theor Appl Genet*, **110(8)**, **1490-1498**.

Ming, R., Liu, S.C., Lin, Y.R., da Silva, J., Wilson, W., Braga, D., van Deynze, A., Wenslaff, T.F., Wu, K.K., Moore, P.H., Burnquist, W., Sorrells, M.E., Irvine, J.E., Paterson, A.H. (1998) Detailed alignment of saccharum and sorghum chromosomes : comparative organization of closely related diploid and polyploid genomes. *Genetics*, **150**, **1663-1682**.

Mitchell-Olds, T., Clauss, M.J. (2002) Plant evolutionary genomics. *Current Opinion in Plant Biology*, **5**, **74-79**.

Mokrani, L., gentzbittel, L., Azanza, F., Fitamant, L., Rachid Al-Chaarani, G., Sarrafi, A. (2002) Mapping and analysis of quantitative trait loci for grain oil content and agronomic traits using AFLP and SSR in sunflower (*Helianthus annuus* L.). *Theor Appl Genet*, **106**, **149-156**.

Mouzeyar, S., Roeckel-Drevet, P., Gentzbittel, L., Philippon, J., Tourvieille de Labrouhe, D., Vear, F., Nicolas, P (1995) RFLP and RAPD mapping of the sunflower *Pl1* locus for resistance to *Plasmopara halstedii* race 1. *Theor Appl Genet*, **91**, **733-737**.

Muller, C., Denis, M., Gentzbittel, L., Faraut, T. (2004) The Iccare web server : an attempt to merge sequence and mapping information for plant and animal species. *Nucleic Acids Res.*, **32**, **W429-W434**.

Mungall, A.J., Palmer, S.A., Sims, S.K., Edwards, C.A., Ashurst, J.L., *et al.* (2003) The DNA sequence and analysis of human chromosome 6. *Nature*, **425**, **805-813**.

N

Nadeau, J.H., Taylor, B.A. (1984) Lengths of chromosomal segments conserved since divergence of man and

mouse. *PNAS*, **81(3)**, **814-818**.

Niwa, Y., Goto, K., Shimizu, M., Kobayashi, H. (1997) Chromosomal Mapping of Genes in the *rbcS* Family in *Arabidopsis thaliana*. *DNA Res.*, **4(5)**, **341-343**.

O

O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E., Marshall Graves, J.A. (1999) The Promise of Comparative Genomics in Mammals. *Science* **286**, **458-481**.

O'Neill, C.M., Bancroft, I. (2000) Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. *The Plant Journal*, **23(2)**, **233-243**.

Oakey, R.J., Watson, M.L., Seldin, M.F. (1992) Construction of a physical map on mouse and human chromosome 1 : comparison of 13 Mb of mouse and 11 Mb of human DNA. *Hum Mol Genet*, **1(8)**, **613-620**.

Oh, K., Hardeman, K., Ivanchenko, M.G., Ellard-Ivey, M., Nebenführ, A., White, T.J., Lomax, T.L. (2002) Fine mapping in tomato using microsynteny with the *Arabidopsis* genome : the *Diageotropica* (Dgt) locus. *Genome Biology*, **3(9)**, **research0049.1-0049.11**.

P

Parkin, I.A.P., Lydiate, D.J., Trick, M. (2002) Assessing the level of collinearity between *Arabidopsis thaliana* and *Brassica napus* for *A. thaliana* chromosome 5. *Genome*, **45**, **356-366**.

Paterson, A.H., Bowers, J.E., Burow, M.D., Draye, X., Elsik, C.G., Jiang, C.-X., Katsar, C.S., lan, T.-H., Lin, Y.-R., Ming, R., Wright, R.J. (2000) Comparative Genomics of Plant Chromosomes. *The Plant Cell*, **12**, **1523-1539**.

Paterson, A.H., Bowers, J.E., Chapmann, B.A. (2004) Ancient polyploidization predating divergence of the cerals, and its consequences for comparative genomics. *PNAS*, **101(26)**, **9903-9908**.

Pereira, M.G., Lee, M., Bramel-Cox, P., Woodman, W., Doebley, J., Whitkus, R. (1994) Construction of an RFLP map in sorghum and comparative mapping in maize. *Genome*, **37**, **236-243**.

Pichersky, E., Bernatzsky, R., Tanksley, S.D., Cashmore, A.R. (1986) Evidence for selection as a mechanism in the conceted evolution of *Lycopersicon esculentum* (tomato) genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase. *PNAS*, **83**, **3880-3884**.

Pontes, O., Neves, N., Silva, M., Lewis, M.S., Madlung, A., Comai, L., Viegas, W., Picaard, C.S. (2004) Chromosomal locus rearrangements are a rapid response to formation of allotetraploid *Arabidopsis suecica* genome. *PNAS*, **101(52)**, **18240-18245**.

Proudhon, D., Wei, J., Briat, J., Theil, EC. (1996) Ferritin gene organization : differences between plants and animals suggest possible kingdom-specific selective constraints. *J Mol Evol*, **42(3)**, **325-336**.

R

Rachid Al-Chaarani, G., Roustaee, A., Gentzbittel, L., Mokrani, L., Barrault, G., Dechamp-Guillaume, G., Sarrafi, A. (2002) A QTL analysis of sunflower partial resistance to downy mildew (*Plasmopara halstedii*) and black stem (*Phoma macdonaldii*) by the use of recombinant inbred lines (RILs). *Theor Appl Genet*, **104**, **490-496**.

Rachid Al-Chaarani, G., Gentzbittel, L., Huang, X.Q., Sarrafi, A. (2004) Genotypic variation and identification of QTLs for agronomic traits, using AFLP and SSR markers in RILs of sunflower (*Helianthus annuus* L.). *Theor Appl Genet*, **109**, **1353-1360**.

Radwan, O., Bouzidi, M.F., Vear, F., Philippon, J., Tourvieille de Labrouhe, D., Nicolas, P., Mouzeyar, S. (2003) Identification of non-TIR-NBS-LRR markers linked to the *Pl5/Pl8* locus for resistance to downy mildew in sunflower. *Theor Appl Genet*, **106**, **1438-1446**.

Radwan, O., Bouzidi, M.F., Nicolas, P., Mouzeyar, S. (2004) Development of PCR markers for the *Pl5/Pl8* locus for resistance to *Plasmopara halstedii* in sunflower, *Helianthus annuus* L. from complete CC-NBS-LRR sequences. *Theor Appl Genet*, **109**, **176-185**.

Radwan, O., Mouzeyar, S. Nicolas, P., Bouzidi, M.F. (2005) Induction of a sunflower CC-NBS-LRR resistance gene analogue during incompatible interaction with *Plasmopara halstedii*. *Journal of Experimental Botany*, **56(412)**, **567-575**.

Raes, J., Vandepoele, K., Simillion, C., Saeys, Y., Van de Peer, Y. (2003) Investigating ancient duplication events in the *Arabidopsis* genome. *Journal of Structural and Functional Genomics*, **3**, **117-129**.

Ramakrishna, W., Dubcovsky, J., Park, Y.J., Busso, C., Emberton, J., SanMiguel, P., Bennetzen, J.L. (2002) Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics*, **162(3)**, **1389-1400**.

Rieseberg, L.H., Linder, C.R., Seiler, G.J. (1995) Chromosomal and Genic Barriers to Introgression in Helianthus. *Genetics*, **141**, **1163-1171**.

Rieseberg, L.H. (2001) Polyploid evolution : Keeping the peace at genomic reunions. *Current Biology*, **11**, **R925-R928**.

Rink, A., Santschi, M.E., Eyer, K.M., Roelofs, B., Hess, M., Godfrey, M., Karajusuf, E.K., Yerle, M., Milan, D., Beatie, C.W. (2002) A first-generation EST RH comparative map of the porcine and human genome. *Mammalian Genome*, **13**, **578-587**.

Robic, A., Seroude, V., Jeon, J.T., Yerle, M., Wasungu, L., Andersson, L., Gellin, J., Mi-

lan, D. (1999) A radiation hybrid map of the RN region in pigs demonstrates conserved gene order compared with the human and mouse genomes. *Mamm. Genome*, **10(6)**, **565-568**.

Robic, A., Jeon, J.T., Amarger, V., Chardon, P., Looft, C., Andersson, L., Gellin, J., Milan, D. (2001) Construction of a high-resolution RH map of the human 2q35 region on TNG panel and comparison with a physical map of the porcine homologous region 15q25. *Mamm. Genome*, **12(5)**, **380-386**.

Rossberg, M., Theres, K., Acarkan, A., Herrero, A., Schmitt, T., Schumacher, K., Schmitz, G., Schmidt, R. (2001) Comparative Sequence Analysis Reveals Extensive Micorcollinearity in the *Lateral Suppressor* Regions of the Tomato, Arabidopsis, and Capsella Genomes. *The Plant Cell*, **13**, **979-988**.

Rozen, S., Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol.*, **132**, **365-386**.

S

Sahrawy, M., Hecht, V., Lopez-Jaramillo, J., Chueca, A., Chartier, Y., Meyer, Y. (1996) Intron Position as an Evolutionary Marker of Thioredoxins and Thioredoxin Domains. *J Mol Evol*, **42**, **422-431**.

Salanoubat, M., Lemcke, K., Rieger, M., Ansorde, W., Unseld, M., *et al.* (2000) Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature*, **408**, **820-822**.

Salse, J., Piégu, B., Cooke, R., Delseny, M. (2002) Synteny between *Arabidopsis thaliana* and rice at the genome level : a tool to identify conservation in the ongoing rice genome sequencing project. *Nucleic Acids Res.*, **30(11)**, **2316-2328**.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., Bennetzen, J.L. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274(5288)**, **765-768**.

Sankoff, D., Nadeau, J.H. (2003) Chromosome rearrangements in evolution : From gene order to genome sequence and back. *Proc Natl Acad Sci U S A.*, **100(20)**, **11188-11189**.

Santini, S., Cavallini, A., Natali, L., Minelli, S., Maggini, F., Cionini, P.G. (2002) *Ty*1/*copia*- and *Ty*3/*gypsy*-like DNA sequences in *Helianthus* species. *Chromosoma*, **111**, **192-200**.

Sasanuma, T., Miyashita, N.T. (1998) Subfamily divergence in the multigene family of ribulose-1,5-bisphosphate carboxylase/oxygenase (*rbcS*) in Triticeae and its relatives. *Genes Genet Syst*, **73**, **297-309**.

Sasanuma, T. (2001) Characterization of the *rbcS* multigene family in wheat : subfamily classification, determination of chromosomal location and evolutionary analysis. *Mol Genet Genomics*, **265**, **161-171**.

Schibler, L. Vaiman, D., Oustry, A., Giraud-Delville, C., Cribiu, E.P. (1998) Comparative gene mapping : a fine-scale survey of chromosome rearrangements between ruminants and humans. *Genome Res.*,

8(9), **901-915**.

Schilling, E.E. (1997) Phylogenetic analysis of Helianthus (Asteraceae) based on chloroplast DNA restriction site data. *Theor Appl Genet*, **94**, **925-933**.

Schmidt, T. (1999) LINEs, SINEs and repetitive DNA : non-LTR retrotransposons in plant genomes. *Plant Mol Biol*, **40**, **903-910**.

Schmidt, T. (2000) Synteny : recent advances and future prospects. *Current Opinion in Plant Biology*, **3**, **97-102**.

Schoof, H., Ernst, R., Nazarov, V., Pfeifer, L., Mewes, H.W., Mayer, K.F. (2004) MIPS Arabidopsis thaliana Database (MatDB) : an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res.*, **32(Database issue)**, **D373-D376**.

Schrader, O., Ahne, R., Fuchs, J., Schubert, I. (1997) Karyotype analysis of *Helianthus annuus* using Giemsa banding and fluorescence *in situ* hybridization. *Chromosome Res.*, **5**, **451-456**.

Schwarzbach, A.E., Donovan, L.A., Rieseberg, L.H. (2001) Transgressive character expression in a hybrid sunflower species. *American Journal of Botany*, **88(2)**, **270-277**.

Seoighe, C., Wolfe, K.H. (1999) Updated map of duplicated regions in the yeast genome. *Gene*, **238(1)**, **253-261**.

Seoighe, C., Wolfe, K.H. (1999) Yeast genome evolution in the post-genome era. *Curr Opin Microbiol*, **2(5)**, **548-554**.

Silverthorne, J., Wimpee, C.F., Yamada, T., Rolfe, S.A., Tobin, E.M. (1990) Differential expression of individual genes encoding the small subunit of ribuloase 1,5-bisphosphate carboxylase in *Lemna gibba*. *Plant Mol Biol*, **15**, **49-58**.

Simillion, C., Vandepoele, K., Van Montagu, M.C.E., Zabeau, M., Van de Peer, Y. (2002) The hidden duplicaiton past of *Arabidopsis thaliana*. *PNAS*, **99(21)**, **13627-13632**.

Song K., Lu P., Tang K., Osborn T.C. (1995) Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploids evolution. *PNAS*, **92**, **7719-7723**.

Song, R., Llaca, V., Messing, J. (2002) Mosaic Organization of Orthologous Sequences in Grass Genomes. *Genome Res.*, **12**, **1549-1555**.

Sossey-Alaoui, K., Serieys, H., Tersac, M., Lambert, P., Schilling, E., Griveau, Y., Kaan, F., Bervillé, A. (1998) Evidence for several genome in *Helianthus*. *Theor Appl Genet*, **97**, **422-430**.

Spreitzer, R.J., Salvucci, M.E. (2002) Rubisco : Struture, Regulatory interactions, and Possibilities for a Better Enzyme. *Annu Rev Plant Biol*, **53**, **449-475**.

Spreitzer, R.J. (2003) Role of the small subunit in ribulose-1,5-biphosphate carboxylase/oxygenase. *Archives of Biochemistry and Biophysics*, **414**, **141-149**.

Stuart-Rogers, C., Flavell, A.J. (2001) The Evolution of *Ty*1-*copia*-like Group Retrotransposons in Gymnosperms. *Mol Biol Evol*, **18(2)**, **155-163**.

Sugita, M., Manzara, T., Pichersky, E., Cashmore, A., Gruissem, W. (1987) Genomic organization, sequence analysis and expression of all five genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from tomato. *Mol Gen Genet.*, **209(2)**, **247-256**.

Suoniemi, A., Tanskanen, J., Schulman, H. (1998) Gypsy-like retrotransposons are widespread in the plant kingdom. *The Plant Journal*, **13(5)**, **699-705**.

T

Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., *et al.* (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*, **408**, **823-826**.

Tang, S., Yu, J.-K., Slabaugh, M.B., Shintani, D.K., Knapp, S.J. (2002) Simple sequence repeat map of the sunflower genome. *Theor Appl Genet*, **105**, **1124-1136**.

Tanksley, S.D., Ganal, M.W., Prince, J.P., de Vicente, M.C., Bonierbale, M.W., Broun, P., Fulton, T.M., Giovannoni, J.J., Grandillo, S., Martin, G.B., Messenguer, R., Miller, J.C, Miller, L., Paterson, A.H., Pineda, O., Röder, M.S., Wing, R.A., Wu, W., Young, N.D. (1992) High Density Molecular Linkage Maps of the Tomato and Potato Genoms. *Genetics*, **132**, **1141-1160**.

Taylor, M.S., Semple, C.A. (2002) Sushi gets serious : the draft genome sequence of the pufferfish *Fugu rubripes*. *Genome Biol*, **3(9)**, **reviews1025**.

Theologis, A., Ecker, J.R., Palm, C.J., Federspiel, N.A., Kaul, S., *et al.* (2000) Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature*, **408**, **816-819**.

Thomas, J.W., Prasad, A.B., Summers, T.J., Lee-Lin, S.-Q., Maduro, V.V.B., Idol, J.R., Ryan, J.F., Thomas, P.J., McDowell, J.C., Green E.D. (2002) Parallel Construction of Orthologous Sequence-Ready Clone Contig Maps in Multiple Species. *Genome Res.*, **12**, **1277-1285**.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, **788-793**.

U

Ungerer, M.C., Baird, S.J.E., Pan, J., Rieseberg, L.H. (1998) Rapid hybrid speciation in wild sunflowers. *PNAS*, **95**, **11757-11762**.

V

van Dodeweerd, A.-M., Hall, C.R., Bent, E.G., Johnson, S.J., Bevan, M.W., Bancroft, I. (1999) Identification and analysis of homoeologous segments of the genomes of rice and *Arabidopsis thaliana*. *Genome*, **42**, **887-892**.

Vandepoele, K., Simillion, C., Van de Peer, Y. (2002) Detecting the undetectable : uncovering duplicated segments in *Arabidopsis* by comparison with rice. *Trends in Genetics*, **18(12)**, **606-608**.

Vear, F., Gentzbittel, L., Philippon, J., Mouzeyar, S., Mestries, E., Roeckel-Drevet, P., Tourvieille de Labrouhe, D., Nicolas, P. (1997) The genetics of resistance to five races of downy mildew (*Plasmopara halstedii*) in sunflower (*Helianthus annnus* L.). *Theor Appl Genet*, **95**, **584-589**.

Ventelon, M., Deu, M., Garsmeur, O., Doligez, A., Ghesquière, A., Lorieux, M., Rami, J.F., Glaszmann, J.C., Grivet, L. (2001) A direct comparison between the gentic maps of sorghum and rice. *Theor. Appl. Genet.*, **102**, **379-386**.

Vision, T.J., Brown, D.G., Tanksley, S.D. (2000) The Origins of Genomic Duplications in *Arabidopsis*. *Science*, **290**, **2114-2117**.

Voytas, D.F., Cummings, M.P., Konieczny, A., Ausubel, F.M., Rodermel, S.R. (1992) *copia*-like retrotransposons are ubiquitous among plants. *PNAS*, **89**, **7124-7128**.

W

Waksman, G., Lebrun, M., Freyssinet, G. (1987) Nucleotide sequence of a gene encoding sunflower ribulose-1,5- bisphosphate carboxylase/oxygenase small subunit (rbcS). *Nucleic Acids Res.*, **15(7)**, **7181**.

Waksman, G., Freyssinet, G. (1987) Nucleotide sequence of a cDNA encoding the ribulose-1,5-bisphosphate carboxylase/oxygenase. *Nucleic Acids Res.*, **15(3)**, **1328**.

Wang, J., Tian, L., Madlung, A., Lee, H.-S., Chen, M., Lee, J.J., Watson, B., Kagochi, T., Comai, L., Chen, J. (2004) Staochastic and Epigenetic Changes of Gene Expression in Arabidopsis Polyploids. *Genetics*, **167**, **1961-1973**.

Wang, X., Shi, X., Hao, B., Ge, S., Luo, J. (2005) Duplication and DNA segmental loss in the rice genome : implications for diploidization. *New Phytologist*, **165**, **937-946**.

Watkins-Chow, D.E., Buckwalter, M.S., Newhouse, M.M., Lossie, A.C., Brinkmeier, M.L., Camper, S.A. (1997) Genetic mapping of 21 genes on mouse chromosome 11 reveals disruptions in linkage conservation with human chromosome 5. *Genomics*, **40(1)**, **114-122**.

Wendel, J. (2000) Genome evolution in polyploids. *Plant Mol Biol*, **42**, **225-249**.

White, S.E., Habera, L.F., Wessler, S.R. (1994) Retrotransposons in the flanking regions of normal plant genes : a role for copia-like elements in the evolution of gene structure and expression. *PNAS*, **91**, **11792-11796**.

Wolfe, K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet*, **2(5)**, **333-341**.

Wolter, F.P., Fritz, C.C., Willmitzer, L., Schell, J., Schreier, P.H. (1988) *rbcS* genes in *Solanum tuberosum* : Conservation of transit peptide and exon shuffling during evolution. *PNAS*, **85**, **846-850**.

Y

Yan, H.H., Mudge, J., Kim, D.-J., Larsen, D., Schoemaker, R.C., Cook, D.R., Young, N.D. (2003) Estimates of conserved microsynteny among the genomes of *Glycine max*, *Medicago truncatula* and *Arabidopsis thaliana*. *Theor Appl Genet*, **106**, **1256-1265**.

Yu, J.-K., Mangor, J., Thompson, L., Edwards, K.J., Slabaugh, M.B., Knapp, S.J. (2002) Allelic diversity of simple sequence repeats among elite inbred lines of cultivated sunflower. *Genome*, **45**, **652-660**.

Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., Liu, B., Deng, Y., *et al.* (2002) A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, **79-92**.

Z

Zhang, Y., Xu, G.-H., Guo, X.-Y., Fan, L.-J (2005) Two ancient rounds of polyploidy in rice genome. *J Zhejiang Univ SCI*, **6B(2)**, **87-90**.

Annexe A

Erreurs de Séquençage

Les séquences EST disponibles dans les bases de données sont issues de la transcription réverse de séquences ARNm. Ces ARNm sont extraits des cellules végétales puis ils subissent une réverse transcription pour obtenir un brin d'ADN dit complémentaire (ADNc). Les simples brins d'ADNc sont ensuite doublés et inserrés dans des vecteurs afin d'être séquencés. Le séquençage s'effectue généralement avec des amorces universelles c'est-à-dire qu'elles s'hybrident sur le vecteur et permettent d'amplifier la séquence ADNc inserrée mais aussi une partie de vecteur.

Une fois séquencées, ces séquences sont analysées en utilisant des outils bioinformatiques afin d'éliminer les séquences de vecteur pour n'avoir que la séquence ADNc inserrée. La séquence obtenue après le séquençage est associée avec deux fichiers contenant la probabilité d'exactitude des nucléotides de la séquence et le profil électrophorétique de la séquence. A partir de ces deux fichiers la séquence de l'ADNc est reconstituée et une probabilité d'exactitude est associée à chaque nucléotide de la séquence. Le début et la fin de séquence, la probabilité d'exactitude des nucléotides est plus faible qu'au milieu de la séquence. De plus, un certain nombre d'erreur d'interprétation de l'électrophorégramme peuvent être faite.

La Figure A.1 illustre les quelques erreurs de séquençage qui peuvent se produire. Quatre séquences codantes du gène d'*Arabidopsis* codant pour la sous-unité S de la Rubisco (At5g38410, At5g38420, At5g38430 et At1g67090) sont alignées avec 9 séquences EST de tournesol. Les séquences codantes (partie exonique traduite du gène) des gènes d'*Arabidopsis* ont été vérifiées expérimentalement et codent bien pour la Rubisco, ce qui implique que le cadre ouvert de lecture est correct. Les séquences de tournesol s'alignent avec les gènes d'*Arabidopsis*. En position 303 à 305, trois nucléotides sont présents chez les gènes d'*Arabidopsis* mais pas sur les séquences EST de tournesol. Ce n'est pas une erreur de séquences car toutes les séquences EST ne possèdent pas ces trois nucléotides, de plus l'absence de trois nucléotides ne perturbe pas le cadre de lecture des séquences de tournesol. En revanche, en position 321, 346, $364-366$, 374, $391-396$, 480, $496-497$ et 510, des nucléotides ne sont présents que sur une des séquences

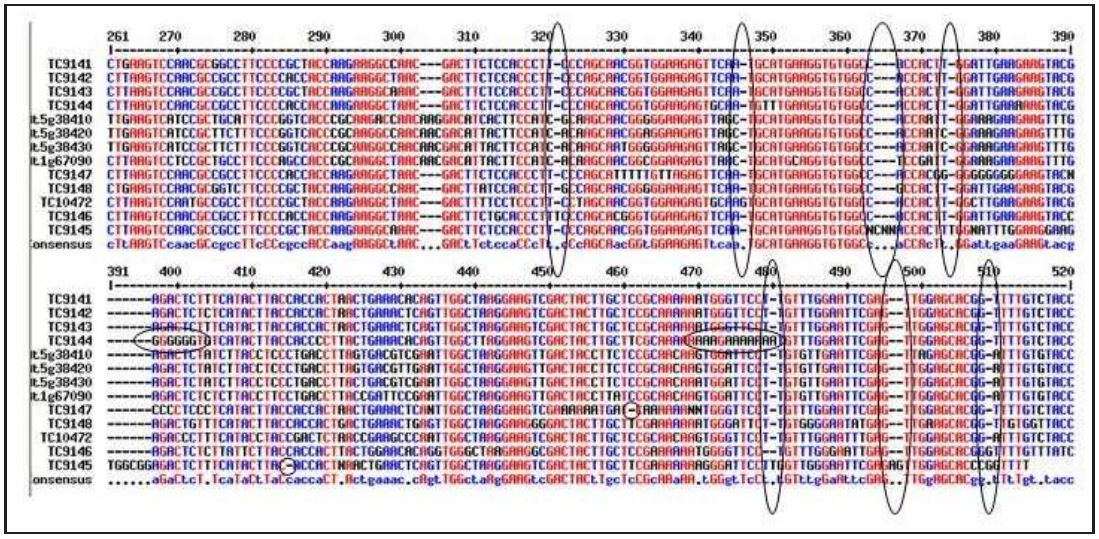


FIG. A.1 – Alignement multiple de 9 séquences EST contre 4 séquences codantes de gènes d'*Arabidopsis* qui illustrent bien les erreurs de séquençage des EST.

EST de tournesol et pas sur les autres séquences EST ni sur les gènes d'*Arabidopsis*. Ces nucléotides surnuméraire perturbent le cadre de lecture et ne sont pas présents sur d'autres séquences de tournresol, il doit donc s'agir d'erreur de séquençage. En position, 415 et 461, il manque un nucléotide dans des séquences EST de tournesol. De même, au environ des bases 400 et 470 de la séquence EST TC9144, des nucléotides totalement différents des autres séquences sont observés, il doit aussi s'agir d'erreurs de séquençage.

Parmi les différentes séquences EST utilisées, certaines d'entre elles contiennent des erreurs de séquençage (insertion, délétion ou substitution de nucléotides).

Annexe B

Du Gène à la Protéine

Paradigme de la biologie moléculaire :

"Les gènes sont des unités qui se perpétuent et qui fonctionnent à travers leur expression sous forme de protéines."

L'information génétique est portée par la séquence de l'ADN. L'information est exprimée grâce à un mécanisme en deux étapes. La **transcription** engendre un ARN simple brin dont la séquence est identique à l'un des deux brins de l'ADN bicaténaire. Puis, la **traduction** convertit la séquence nucléotidique de l'ARN en une séquence d'acides aminés constituant une protéine.

La Figure B.1 présente les différents processus qui permettent l'expression d'un gène et les différents protagonistes qui interviennent. Un gène est localisé sur l'ADN génomique. Il est constitué d'une partie promotrice en amont (en bleu) qui contrôle son expression, puis d'une alternance d'exons (en rouge) et d'introns (en vert). L'expression de ce gène débute par la **transcription**. Celle-ci commence par la réplication du brin codant de l'ADN sous forme simple brin. L'ADN du brin codant du gène sert de matrice au futur ARN messager (ou ARNm). Cette réplication aboutit à la synthèse d'une séquence nucléotidique simple brin appelée ARN primaire ou pré-ARNm où les nucléotides Thymidines "T" sont remplacés par des nucléotides Uraciles "U". Cet ARN primaire est constitué de l'ensemble des exons et des introns du gène. Ensuite, cet ARN primaire subit une maturation. Dans la partie 5' terminale, la guanylyl transférase ajoute un nucléotide de Guanidine par un pont 5'-5' triphosphates. La partie terminale 3' se voie ajouter une queue polyA (jusqu'à 200 résidus A). La séquence ARN primaire subit l'excision des introns et l'épissage des exons de manière à obtenir un ARN messager mature (ARNm). L'ARNm est constitué uniquement des séquences exoniques.

Après la transcription intervient la **traduction** de cet ARNm. Lors de la traduction, l'ARNm va être lu par triplet de nucléotides (formant ainsi un codon) et à chaque codon est associé un acide aminé. La traduction va débuter au codon START (AUG), chaque codon est remplacé par un acide aminé et la traduction se termine au codon STOP (UAA, UAG

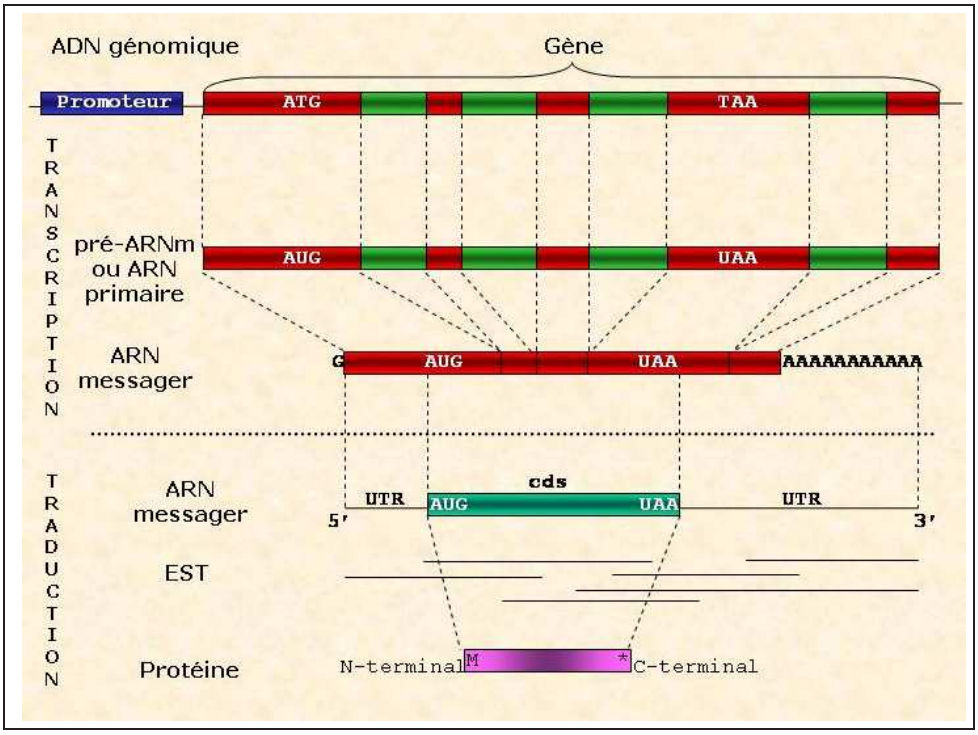


FIG. B.1 – Représentation schématique de la transcription et de la traduction d'un gène.

ou UGA) de l'ARNm. La partie de l'ARNm se trouvant en amont du codon START est la partie appelé UTR 5' (UnTranslated Region 5' pour Région 5' non traduite) et celle en aval du codon STOP est la partie UTR 3'. L'ARNm est donc constitué de trois parties : une partie UTR 5', une région dite codante (ou région cds) entre le codon START et le codon STOP et une partie UTR 3'. Seule la région codante servira de matrice à la synthèse de la protéine, les deux parties UTR ne sont pas traduites.

Les séquences EST sont des fragments partiels d'ARNm. Les séquences EST disponibles dans les bases de données sont issues de la transcription réverse à partir d'ARNm. Ces ARNm sont extraits des cellules végétales puis subissent une réverse transcription pour obtenir un brin d'ADN dit complémentaire (ADNc). Les simples brins d'ADN sont ensuite doublés et inserrés dans des vecteurs afin d'être séquencés. les ARNm qui peuvent avoir des tailles très variable, de 200 pb à 4000 pb, ne sont généralement pas séquencés en totalité. La plupart du temps ces ARNm sont séquencés à leur extrémité 3' ou 5'. Les séquences EST sont donc des fragements partiels d'ARNm et un ensemble de séquences EST, aussi appelé TC pour Tentative Consensus, peut permettre de reconstituer la totalité de la séquence de l'ARNm. Il n'est donc pas étonnant que plusieurs séquences EST soient redondantes qu'un ARNm.

Quand deux gènes sont orthologues ou paralogues, le degré d'homologie est moins important dans les introns que dans les exons. Etant donné que les exons doivent coder pour des séquences d'acides aminés, leur potentiel de changement de séquence est limité. Libres de cette contrainte, les introns peuvent subir et accumuler plus de substitutions que les exons. De

même, pour les régions UTR 3' et 5'. Le degré de divergence entre deux exons est corrélé aux différences existant entre les protéines. Il est généralement le résultat de substitutions de bases. Cependant, les modifications qui n'affectent pas la siginification des codons sont nombreuses, soit parce qu'elles touchent la position de la troisième base du codon, soit parce qu'elles sont situées dans des régions non traduites. Les introns évoluent beaucoup plus rapidement que les exons ; si l'on compare le même gène dans des espèces différentes, les exons sont quelquefois homologues alors que les introns ont tant divergé que les séquences équivalentes ne peuvent être reconnues. Les mutations apparaissent à des taux identiques dans les exons et dans les introns, mais disparaissent plus efficacement dans les exons par contre-sélection.

En définitive lorsque l'on compare des gènes, seules les séquences exoniques présenteront des similitudes. Ceci implique que les séquences EST, qui sont donc des fragments partiels d'ARNm, ont des séquences nucléotidiques fortement conservées pour la région codante (qui correspond aux exons) et beaucoup plus variable dans les régions UTR3' et UTR5' pour un même gène.

Annexe C

The Iccare Web Server

Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W429–W434***
DOI: 10.1093/nar/gkh460

The Iccare web server: an attempt to merge sequence and mapping information for plant and animal species

Cédric Muller[1], Mathieu Denis, Laurent Gentzbittel[1] and Thomas Faraut*

INRA, Laboratoire de génétique cellulaire and [1]INP-ENSAT, Laboratoire de biotechnologies et d'amélioration des plantes, Castanet Tolosan 31326, France

Received February 16, 2004; Revised and Accepted April 26, 2004

ABSTRACT

The Iccare web server, http://genopole.toulouse.inra.fr/bioinfo/Iccare, provides a simple yet efficient tool for crude EST (expressed sequence tag) annotation specifically dedicated to comparative mapping approaches. Iccare uses all the EST and mRNA sequences from public databases for an organism of interest (query species) and compares them to all the transcripts of one reference organism (*Homo sapiens* or *Arabidopsis thaliana*). The results are displayed according to the location of the genes on the chromosomes of the reference organism. Gene structure information and sequence similarities are combined in a graphical representation in order to pinpoint the nature of the transcript query sequence. The user can subsequently design primers or probes for the purpose of physical or genetic mapping. In addition to the query organisms already available in Iccare, users can perform a tailor-made search with their own sequences against the animal or plant reference organism genes.

INTRODUCTION

Comparative analysis has always been central to biology. With the advent of the DNA revolution, the development of computer tools for sequence comparison has been essential to apply a comparative approach at the protein or DNA level. The success of this approach is attested by the extensive use of the world-famous BLAST programs [(1), W. Gish, http://blast.wustl.edu]. In the field of gene mapping, the comparative approach takes advantage of genome conservation. Indeed, while at the molecular level genes evolve by accumulated point mutations, at the genome scale chromosomes evolve by inter- and intra-chromosomal rearrangements of large segments within which the gene content and order may remain unaltered (2–4). Comparative mapping studies have confirmed the local conservation of gene repertoire and order, also called microsynteny conservation, not only between mammals (5–7) but also between mammalian species and more distant animals such as pufferfish and chicken (8–10). For plants, this conservation is clearly revealed between species of the same family [Poaceae (11–13) or Brassicaceae (14–16)] and also between *Arabidopsis thaliana* (Brassicaceae) and species of other families [Fabaceae (17–19), Poaceae (20,21) and Solanaceae (22,23)]. Without underestimating the confusing effect of local micro-rearrangements, the synteny conservation enables, at least to a certain extent, the transfer of mapping information from one species to another and many genome species are now studied using the human (24–28) or the *A.thaliana* (29–33) genome dense maps to accelerate the process of mapping.

With the emergence of high-throughput sequencing projects, which generate both EST (expressed sequence tag) and genomic DNA, an impressive amount of sequence data from many species is available in public databases. Whereas much effort has been devoted to developing computer tools for sequence assembly, annotation (34,35) and their graphical display (36,37), the process of gathering sequence data in a chromosomal region of interest for a given genome is still a time-consuming and awkward task. The main comparative mapping studies are usually oriented towards local chromosomal investigations for Quantitative Trait Loci (QTL) mapping or fine mapping purposes. In the case of a fine mapping approach in a mammalian species, the process starts by identifying, using state-of-the-art comparative maps, the corresponding region in the human genome. The gene sequences located in the so-called homologous region are used to identify available orthologous ESTs for the genome under study. These EST sequences are subsequently used to design primers for mapping experiments.

To facilitate and accelerate the exploitation of public sequence data available for different plant and animal species—the query species—we propose organizing them according to their sequence similarities to the genes of a completely sequenced organism—reference organism (*Homo sapiens* for animals and *A.thaliana* for plants). This is the underlying

*To whom correspondence should be addressed. Tel: +33 05 61 28 54 31; Fax: +33 0 5 61 28 53 08; Email: Thomas.Faraut@toulouse.inra.fr

principle of the tool Iccare, which stands for 'Interspecific Comparative Clustering and Annotation foR Est', available at http://genopole.toulouse.inra.fr/bioinfo/Iccare. Using this tool, sequence similarities results are available through the chromosomal maps of the reference organism. Furthermore, based on the gene splicing information for the reference organism, Iccare enables the prediction of splicing site positions on the query species' sequences. In addition to publicly available data, the user can organize new sequences according to the same scheme.

MATERIALS AND METHODS

Datasets

For the human genome, the UniGene clusters were used to define the reference organism gene catalog [(38), http://www.ncbi.nlm.nih.gov]. The set of unique sequences is used to define the reference transcript sequence for each gene (see the README file at ftp://ftp.ncbi.nih.gov/repository/UniGene). The mapping information as well as the gene structure information is defined according to the annotation of the human genome provided by Ensembl [(37), http://www.ensembl.org]. For the *Arabidopsis* genome, the gene sequences (26, 637), gene information and mapping information were defined according to the Munich Information Center for Protein Sequences [(39), http://mips.gsf.de]. For *Arabidopsis* transcript sequences, only the translated regions of the predicted mRNA were used. For the query animal or plant species (Table 1), all EST and full-length mRNA sequences have been downloaded using the SRS retrieval software on the Infobiogen server (http://www.infobiogen.fr). Additional organisms can be added on request to the authors.

Personal inputs

Personal sequences can be also submitted to Iccare. The input required by Iccare is a sequence or a set of sequences in FASTA format.

Software description

All sequences are screened for vectors and masked for known repeats with RepeatMasker (UniVec at ftp://ftp.ncbi.nih.gov/pub/UniVec; A. Smit and P. Green, RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html). The masked sequences are subsequently compared to all the transcript sequences of the reference organism with the blastn option of the BLAST program (1). The results are filtered according to the expected value. This value is normalized in order to fit a standardized expected (*E*-)value of a comparison with a database of one million residues. Only similarities—or more precisely, using the terminology of BLAST, highest scoring pairs (HSPs)—with an *E*-value $<10^{-5}$ are kept for further consideration (the complete BLAST output is also available). Finally, Iccare compiles the BLAST results and the mapping information for the reference organism and formats these results for the web site application.

Website dynamic script

Programming was done in Perl using the CGI library, the GD library and the EMBOSS package (40).

Table 1. Available organisms on Iccare

Animals	Plants
Bos Taurus	*Brassica napus*
Canis familliaris	*Brassica rapa*
Capra hircus	*Chlamydomonas reinhardtii*
Equus caballus	*Helianthus annuus*
Gallus gallus	*Medicago truncatula*
Oryctolagus cuniculus	*Physcomitrella patens*
Ovis aries	*Oryza sativa*
	Zea mays

ICCARE WEBSITE

The taxonomic tree on the Iccare homepage presents the different query organisms that can be selected for analysis. These organisms are compared to an appropriate reference organism (currently limited to *H.sapiens* for animal species and *A.thaliana* for plant species). The results are presented on dynamic web pages. A tutorial and help pages are also available.

After having selected the query organism and a model chromosome, the first result page (Figure 1) presents a graphical representation of the chromosome of the reference organism together with a visualization of the gene distribution with particular emphasis on genes showing sequence similarities to the sequences of the query organism. Moreover, a BLAST link allows a blatsn or tblastx search for each query organism sequence against the gene catalog of the reference organism. This gives an idea of the gene family potentially associated with the putative ortholog. The 'alignment' link provides a graphical representation of the gene structure of the reference organism (translated region, exon splicing positions and intron size) and the query sequence similarity with this gene (Figure 2A). The alignment—global or local—between the reference and the query sequences is also available (Figure 2B). As the gene structure is known to be very well conserved (41), this information makes it possible to predict the structure (exon borders) of the query sequence. Various links have been devised for the sake of practicality: Primer3 for primer design (42) and Overgo Maker 40 (http://www.genome.wustl.edu).

RESULTS AND DISCUSSION

The Iccare tool has proven to be practical and efficient for various animal and plant studies. In animals, Iccare has been used in the context of comparative mapping studies between human and pig (43–45), chicken (46) and bovine (47). More recently, in plants, the use of Iccare has been of great help for the construction of sunflower genetic maps or physical maps. The possibility of inferring the putative exon splicing sites increases the success rate for designing polymorphic markers from 15–20% by random primer design to 65% (Delphine Samson, GenoPlante, personal communication). In addition, the identification of conserved EST regions facilitates the design of overgos for physical mapping.

Indeed, with the availability of state-of-the-art comparative maps for bovine and human species, it is straightforward to identify all the bovine transcribed sequences available in the public databases that have sequence similarities to the human genes in the region of interest. Moreover, the graphical display

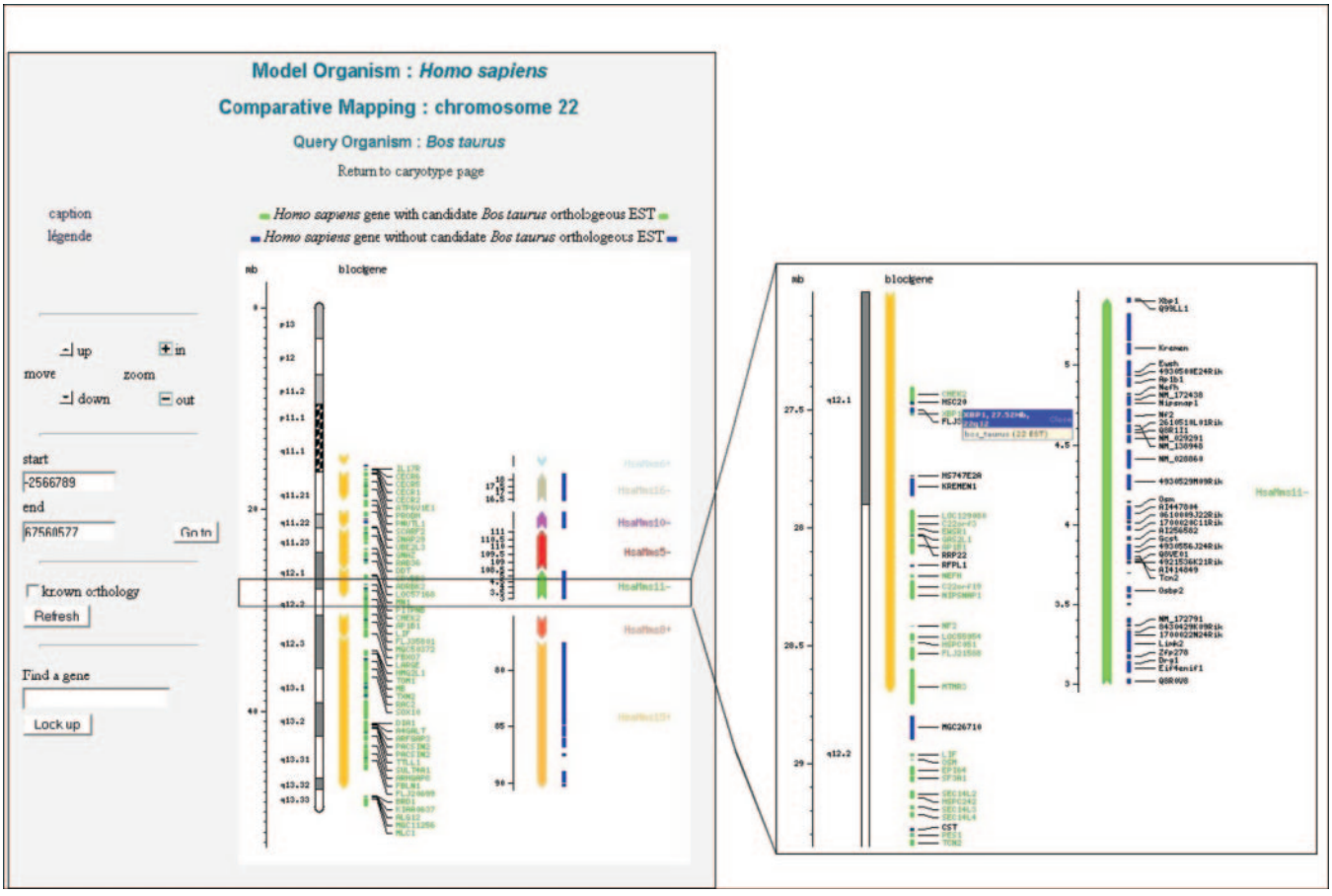


Figure 1. Global chromosome synteny. Example of human chromosome 22 gene map for the query species *Bos taurus* at the chromosome scale on the left and at a higher resolution on the right. From left to right, the human chromosome image displays map locations in Mb and a cytogenetic map with alternating dark and light bands, with a dark-and-white check pattern for the centromeric region. The dark-yellow arrowed rectangles display the conserved syntenic segments shared with the mouse genome. To the right, the corresponding conserved syntenic segments of the mouse genome are displayed in colors corresponding to the mouse chromosome. The buttons on the left enable the user to move to different regions at different levels of resolution. The genes colored in green correspond to human genes for which sequence similarities have been observed to bovine sequences. Human genes without significant similarities to bovine sequences are shown in blue. The mouseover on a gene provides a link that gives access to detailed information about the alignments (see Figure 2).

of sequence similarities allows the addition of pertinent information to the sole information of the *E*-value associated with the score as given by alignment software. A local similarity restricted to the coding region of the human transcript is in good agreement with the scheme of sequence evolution. Furthermore, additional features of the alignment such as phase conservation and mismatches occurring essentially on the third codon position correspond to qualitative characterization of the alignment that are not taken into account by the statistical model of sequence similarities. Finally, additional sequence comparisons can be carried out by the user in order to clarify the relationship between the transcripts of the species of interest and the reference organism transcripts. This might pinpoint for instance an alternative nature of a particular EST.

It has to be pointed out that the distinction between orthologous and paralogous genes on the basis of sequence similarities is a difficult issue. Other programs are specifically dedicated to this purpose (48–50). This problem should therefore be kept in mind when using the simple sequence similarities results provided by Iccare. The question is even more problematic in plants where the genome seems to have evolved by several rounds of polyploidy and/or chromosomal duplication (51). Nevertheless, the simple procedure proposed by our software makes it possible to assume membership of the query sequence to a gene family, which enables further consideration using specific software.

CONCLUSION

With the accumulation of massive amounts of sequence data for many organisms, the problem arises of proposing efficient computer tools to facilitate access to data. A trade-off has to be made between general-purpose databases where all the information is available but difficult to exploit and completely automated systems proposing annotated EST sequences clustered and associated with consensus sequences which are tentatively the reconstructed mRNA. Although both systems are necessary for different purposes, semi-automated systems, such as Iccare, can be of great help for the exploitation of sequence data.

Figure 2. Gene structure representation and local alignment. (**A**) The *Arabidopsis* gene structure of At4g03280 and *Helianthus annuus* sequences with homologous regions with this gene are shown. The top graphical box contains gene structure information related to the *A.thaliana* gene—intron/exon structure of genomic DNA. The green boxes represent the exons, which are concatenated on the second line, mimicking the transcription process. The numbers on top of the first line correspond to intron size. The third line recalls the transcript sequence with a particular emphasis on the translated region, shown in dark blue. The bottom graphical box contains *H.annuus* sequences and representation of local similarities to the *Arabidopsis* gene At4g03280; information for each query sequence is located to the left (GenBank ID, sequence size and BLASTN *E*-value) and the sequence representation is on the right: the black line symbolizes the sequence. Green boxes correspond to sequence similarities with the *Arabidopsis* gene on the same strand, while yellow boxes correspond to similarities in a reverse-complement manner. Sequence similarities to additional query species are also available ('plus' buttons). (**B**) The display associated with the result of a local alignment between *Arabidopsis* gene At4g03280 and the *H.annuus* EST sequence CD847222. BLASTN information (score, *E*-value, identities, gaps and strand) is placed on the top. The query nucleic sequence is represented in black and the *Arabidopsis* nucleic sequence is represented in alternate blue and mauve letters corresponding to the different exons.

ACKNOWLEDGEMENTS

We thank the many researchers who contributed to Iccare by their useful suggestions. We also thank the Génopôle Toulouse Midi-Pyrénées (France), and especially David Allouche, for providing the computer and computer administration resources for data processing and the web service.

REFERENCES

1. Altschul,S.F., Madden,T.L., Schaeffer,A.A., Zhang,J., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
2. Bancroft,I. (2000) Insights into the structural and functional evolution of plant genomes afforded by the nucleotides of chromosomes 2 and 4 of *Arabidopsis thaliana*. *Yeast*, **17**, 1–5.
3. Eichler,E.E. and Sankoff,D. (2003) Structural dynamics of eukaryotic chromosome evolution. *Science*, **301**, 793–797.
4. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.
5. Goureau,A., Garrigues,A., Tosser-Klopp,G., Lahbib-Mansais,Y., Chardon,P. and Yerle,M. (2001) Conserved synteny and gene order difference between human chromosome 12 and pig chromosome 5. *Cytogenet. Cell Genet*., **94**, 49–54.
6. Lopez-Corrales,N.L., Sonstegard,T.S. and Smith,T.P. (1998) Comparative gene mapping: cytogenetic localization of PROC, EN1, ALPI, TNP1, and IL1B in cattle and sheep reveals a conserved rearrangement relative to the human genome. *Cytogenet. Cell Genet*., **83**, 35–38.
7. Martins-Wess,F., Voss-Nemitz,R., Drogemuller,C., Brenig,B. and Leeb,T. (2002) Construction of a 1.2-Mb BAC/PAC contig of the porcine gene RYR1 region on SSC 6q1.2 and comparative analysis with HSA 19q13.13. *Genomics*, **80**, 416–422.
8. Elgar,G., Sandford,R., Aparicio,S., Macrae,A., Venkatesh,B. and Brenner,S. (1996) Small is beautiful: comparative genomics with the pufferfish (*Fugu rubripes*). *T.I.G*., **12**, 145–150.
9. McLysaght,A., Enright,A.J., Skrabanek,L. and Wolfe,K.H. (2000) Estimation of synteny conservation and genome compaction between puferfish (Fugu) and human. *Yeast*, **17**, 22–36.
10. Thomas,J.W., Touchman,J.W., Blakesley,R.W., Bouffard,G.G., Beckstrom-Sternberg,S.M., Margulies,E.H., Blanchette,M., Siepel,A.C., Thomas,P.J., McDowell,J.C. *et al*. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.
11. Feuillet,C. and Keller,B. (1999) High gene density is conserved at syntenic loci of small and large grass genomes. *Proc. Natl Acad. Sci., USA*, **96**, 8265–8270.
12. Bennetzen,J.L. and Ramakrishna,W. (2002) Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Mol. Biol*., **48**, 821–827.
13. Song,R., Llaca,V. and Messing,J. (2002) Mosaic organisation of orthologous sequences in grass genomes. *Genome Res*., **12**, 1549–1555.
14. Acarkan,A., Roßberg,M., Koch,M. and Schmidt,R. (2000) Comparative genome analysis reveals extensive conservation of genome organisation for *Arabidopsis thaliana* and *Capsella rubella*. *Plant J*., **23**, 55–62.
15. O'Neill,C.M. and Bancroft,I. (2000) Comparative physical mapping of segments of genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. *Plant J*., **23**, 233–243.
16. Parkin,I.A.P., Lydiate,D.J. and Trick,M. (2002) Assessing the level of colinearity between *Arabidopsis thaliana* and *Brassica napus* for *A.thaliana* chromosome 5. *Genome*, **45**, 356–366.
17. Grant,D., Cregan,P. and Shoemaker,C. (2000) Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *Proc. Natl Acad. Sci., USA*, **97**, 4168–4173.
18. Foster-Hartnett,D., Mudge,J, Larsen,D., Danesh,D., Yan,H., Denny,R., Peñuela,S. and Young,N.D. (2002) Comparative genomic analysis of sequences sampled from a small region on soybean (*Glycine max*) molecular linkage group G. *Genome*, **45**, 634–645.
19. Yan,H.H., Mudge,J., Kim,D.J., Larsen,D., Schoemaker,R.C., Cook,D.R. and Young,N.D. (2003) Estimates of conserved microsynteny among the genomes of *Glycine max*, *Medicago truncatula* and *Arabidopsis thaliana*. *Theor. Appl. Genet*., **106**, 1256–1265.
20. Mayer,K., Murphy,G., Tarchini,R., Wambutt,R., Volckaert,G., Pohl,T., Düsterhöft,A., Stiekema,W., Entian,K.-D., Terryn,N. *et al*. (2001) Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana*. *Genome Res*., **11**, 1167–1174.
21. Salse,J., Piégu,B., Cooke,R. and Delseny,M. (2002) Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. *Nucleic Acids Res*., **30**, 2316–2328.
22. Ku,H.-M., Vision,T., Liu,J. and Tanksley,S.D. (2000) Comparing sequenced segements of tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl Acad. Sci., USA*, **97**, 9121–9126.
23. Gebhardt,C., Walkemeier,B., Henselewski,H., Barakat,A., Delseny,M. and Stüber,K. (2003) Comparative mapping between potato (*Solanum tuberosum*) and *Arabidopsis thaliana* reveals structurally conserved domains and ancient duplications in the potato genome. *Plant J*., **34**, 529–541.
24. Dunham,I., Shimizu,N., Roe,B.A., Chissoe,S., Hunt,A.R., Collins,J.E., Bruskiewick,R., Beare,D.M., Clamp,M., Smink,L.I. *et al*. (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–496.
25. Chromosome 21 mapping and sequencing consortium (2000) The DNA sequence of human chromosome 21. *Nature*, **405**, 311–319.
26. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
27. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
28. Mungali,A.J., Palmer,S.A., Sims,S.K., Edwards,C.A., Ashurst,J.L., Wilming,L., Jones,M.C., Horton,R., Hunt,S.E., Scott,C.E. *et al*. (2003) The DNA sequence and analysis of human chromosome 6. *Nature*, **425**, 805–812.
29. Lin,X., Kaul,S., Rounsley,S., Shea,T.P., Benito,M.-I., Town,C.D., Fujii,C.Y., Mason,T., Bowman,C.L., Barnstead,M. *et al*. (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*, **402**, 761–768.
30. Mayer,K., Schüller,C., Wambutt,R., Murphy,G., Volckaert,G., Pohl,T., Düsterhöft,A., Stiekema,W., Entian,K.-D., Terryn,N. *et al*. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, **402**, 769–777.
31. Salanoubat,M., Lemcke,K., Rieger,M., Ansorge,W., Unseld,M., Fartmann,B., Valle,G., Blöcker,H., Perez-Alonso,M., Obermaier,B. *et al*. (2000) Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 820–822.
32. Tabata,S., Kaneko,T., Nakamura,Y., Kotani,T., Kato,T., Asamizu,E., Miyajima,N., Sasamoto,S., Kimura,T., Hosouchi,T. *et al*. (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 823–826.
33. Theologis,A., Ecker,J.R., Palm,C.J., Federspiel,N.A., Kaul,S., White,O., Alonso,S.Y., Altafi,H., Araujo,R., Bowmann,C.L. *et al*. (2000) Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 816–819.
34. Chain,P., Kurtz,S., Ohlebush,E. and Slevak,T. (2003) An applications-focused review of comparative genomics tools: capabilities, limitations, and future challenges. *Brief. Bioinformatics*, **4**, 1–20.
35. Mullikin,J.C. and Ning,Z. (2003) The phusion assembler. *Genome Res*., **13**, 81–90.
36. Couronne,O., Poliakov,A., Bray,N., Ishkhanov,T., Ryaboy,D., Rubin,E., Pachter,L. and Dubchak,I. (2003) Startegies and tools for whole-genome alignments. *Genome Res*., **13**, 73–80.
37. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coales,G., Cox,T., Cuff,J. *et al*. (2004) Ensembl 2004. *Nucleic Acids Res*., **32**, D468–D470.
38. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatuva,T.A. *et al*. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res*., **31**, 28–33.
39. Schoof,H., Ernst,R., Nazarov,V., Pfeifer,L., Mewes,H.W. and Mayer,K.F. (2004) MIPS *Arabidopsis thaliana* Database (MatDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res*., **32**, D373–D376.

40. Rice,P., Longden,I. and Bleeasby,A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.
41. Betts,M.J., Guigo,R., Agarwal,P. and Russell,R.B. (2001) Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? *EMBO J*., **20**, 5354–5360.
42. Rozen,S. and Skaletsky,H.J. (2000) Primer3 on WWW for general users and for biologist programmers. In Krawetz,S., Misemer,S. (ed.), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp. 365–386.
43. Demeure,O., Renard,C., Yerle,M., Faraut,T., Riquet,J., Robic,A., Schiex,T., Rink,A. and Milan,D. (2003) Rearranged gene order between pig and human in a QTL region on SSC 7. *Mamm. Genome*, **14**, 71–80.
44. Robic,A., Faraut,T., Iannuccelli,N., Lahbib-Mansais,Y., Cantegrel,V., Alexander,L. and Milan,D. (2003) A new contribution to the integration of human and porcine genome maps: 623 new points of homology. *Cytogenet. Genome Res*., **102**, 100–108.
45. Bosak,N., Faraut,T., Mikawa,S., Uenishi,H., Kiuchi,S., Hiraiwa,H., Hayashi,T. and Yasue,H. (2003) Construction of a high-resolution comparative gene map between swine chromosome region 6q11–>q21 and human chromosome 19 q-arm by RH mapping of 51 genes. *Cytogenet. Genome Res*., **102**, 109–115.
46. Morisson,M., Jiguet-Jiglaire,C., Lemiere,A., Leroux,S., Faraut,T., Yerle,M. and Vignal,A. (2003) A radiation hybrid panel and its use in developing a gene map of the chicken. *Br. Poult. Sci*., **44**, 797–798.
47. Hayes,H., Elduque,C., Gautier,M., Schibler,L., Cribiu,E. and Eggen,A. (2003) Mapping of 195 genes in cattle and updated comparative map with man, mouse, rat and pig. *Cytogenet. Genome Res*., **102**, 16–24.
48. Li,L., Stoeckert,C.J.,Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*., **13**, 2178–2189.
49. Cannon,S.B. and Young,N.D. (2003) OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics*, **4**, 35.
50. Leveugle,M., Prat,K., Perrier,N., Birnbaum,D. and Coulier,F. (2003) ParaDB: a tool for paralogy mapping in vertebrate genomes. *Nucleic Acids Res*., **31**, 63–67.
51. Wolfe,K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nature Rev*., **2**, 333–341.

Annexe D

Multalin pour Iccare

Prenons l'exemple du gène At5g15550 d'*Arabidopsis* et des séquences EST qui lui sont similaires chez le tournesol et la luzerne. Alignons-les avec les deux versions de Multalin (version originale à gauche de la Figure D.1 et version modifiée à droite). Un total de 10 séquences est aligné pour faciliter les calculs de pourcentage. Les 2 versions présentent le même alignement entre les séquences (seul l'ordre dans lequel sont présentées les séquences diffère). L'alignement des séquences ne diffèrent pas car la version de Multalin couplé à Iccare utilise le même algorythme de calcul, seul la visualisation des résultats est différente. Quelle que soit la version utilisée, 6 séquences EST (CB894463, AW685598, CD854754, BF648867 et AJ500072) s'alignent avec la partie 5' du gène d'*Arabidopsis* et 3 séquences (BU017956, BU034179 et CD858257) s'alignent avec la partie 3' du gène.

Dans la version d'origine de Multalin, aucune région n'est conservée à plus de 90% (en rouge) par contre une région est bien conservée à plus de 50% (en bleu) en position $201-578$. Dans cette région, seules 7 séquences (1 gènes d'*Arabidopsis* et 6 séquences de tournesol et luzerne) sont représentées car BU017956, BU034179 et CD858257 ne s'alignent que plus tard (après la position 940). En regardant plus attentivement l'alignement, on s'aperçoit que les 3 premières séquences sur la figure (BU017956, BU034179 et CD858257) ne commencent qu'à environ $900-1000$ bp par rapport à la séquence codante du gène At5g15550. Et ces séquences sont similaires à la séquence d'*Arabidopsis*.

La version d'origine de Multalin calcule le pourcentage sur l'ensemble des séquences alors qu'en réalité avec des séquences EST comparées à des séquences codantes de gènes, 3 séquences ne débutent pas en 5'. Le pourcentage devrait donc être calculé sur 7 séquences et non sur 10. De même dans la partie 3' de la séquence codante où il ne reste plus que 4 séquences, le pourcentage devrait donc être calculé sur 4 et non sur 10.

La version modifiée de Multalin pour Iccare permet de recalculer le pourcentage de similitudes en fonction du nombre de séquences présentes dans la région et non sur la totalité

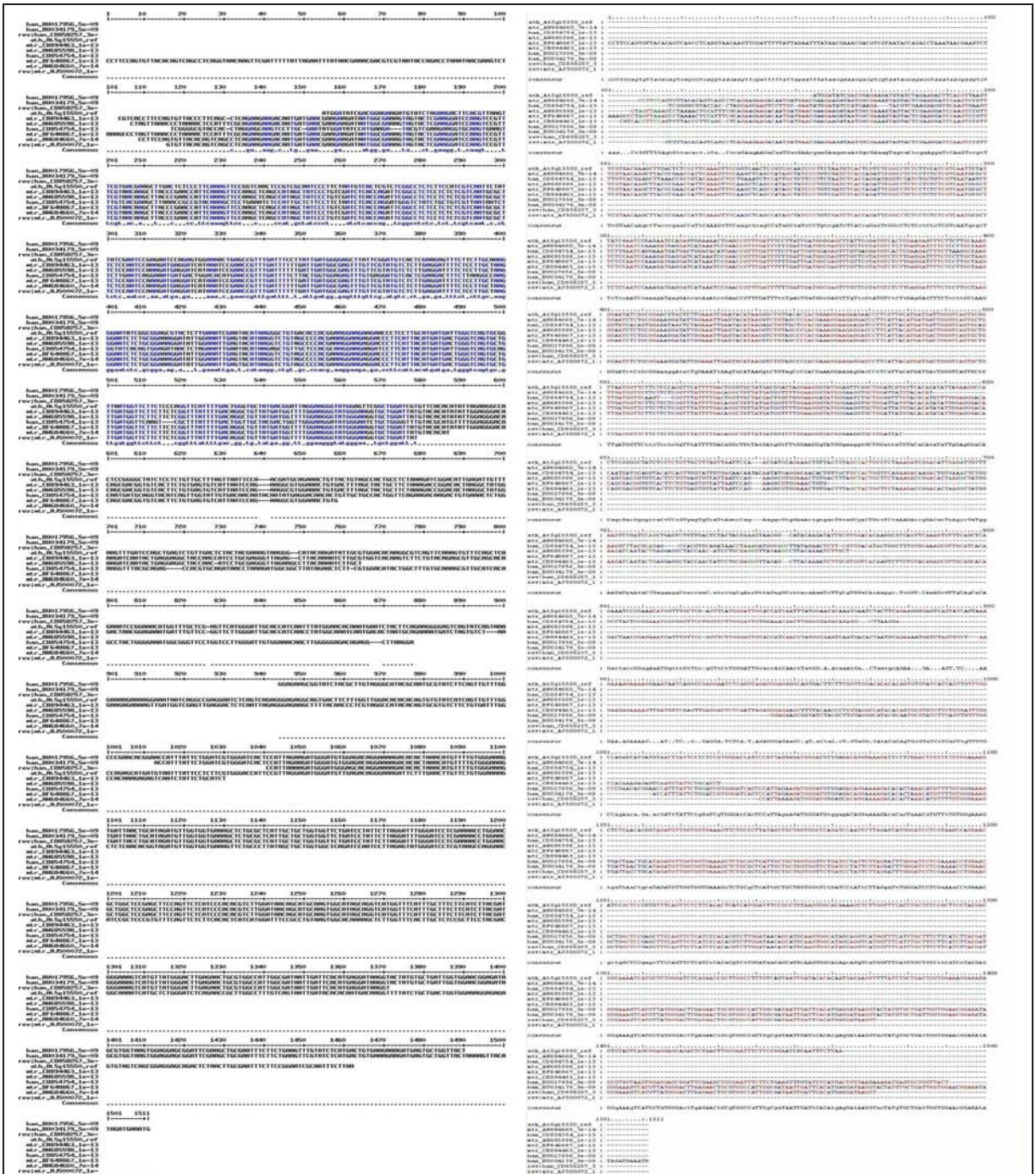


FIG. D.1 – Alignement multiple de dix séquences avec Multalin originale à gauche et Multalin modifiée à droite.

Annexe E

Filtres Haute Densité de la Banque de Clones BAC

La banque de clones BAC est constituée de 412 plaques *GENETIX* au format 384. Ces plaques sont déposées sur des membranes afin de pouvoir procéder au criblage de celle-ci. Les clones BAC sont déposés suivant un pattern de dépôt qui permettra de retrouver l'identité des clones après l'hybridation.

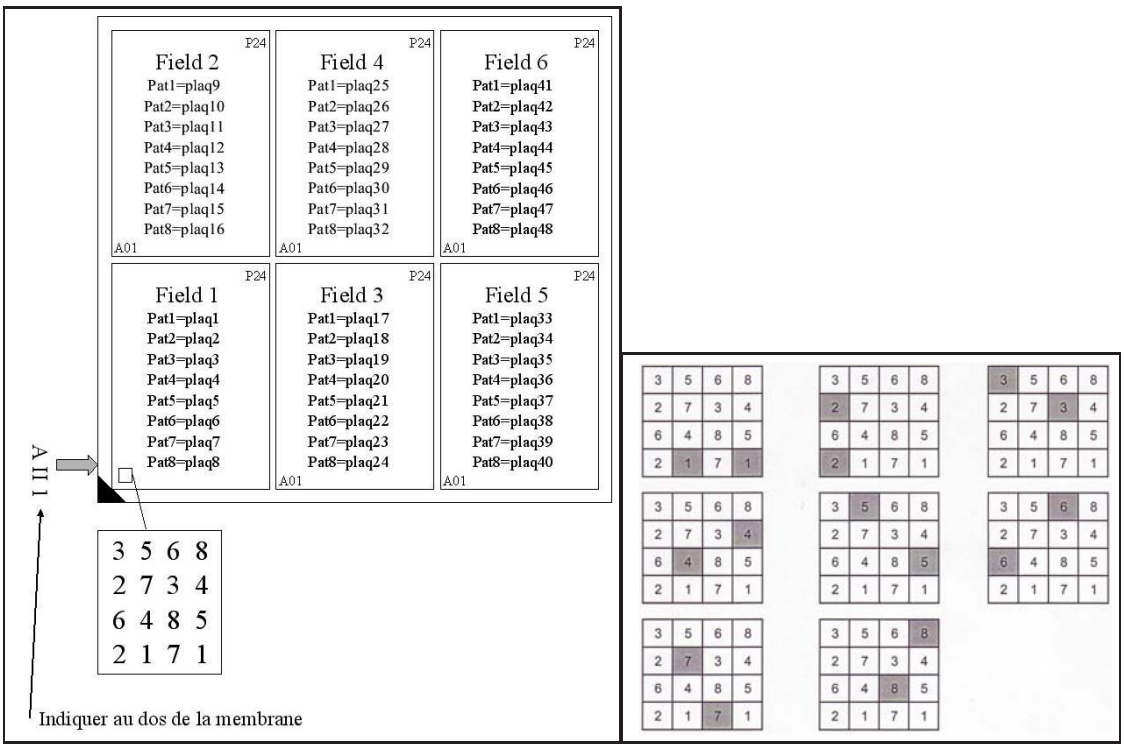


FIG. E.1 – Pattern de dépôt en double dépôt 4x4 des clones BAC sur les membranes.

Les membranes sont divisées en 6 champs pouvant contenir 8 plaques par champ. Chaque

champ est constitué de 384 zones de dépôt (correspondant au puits d'un plaque 384). Ces zones de dépôts contiennent 16 dépôts de clones BAC (dépôt en 4x4). Chaque clone BAC est déposé 2 fois dans une même zone de dépôt suivant un pattern, soit 8 clones BAC différents dans une zone de dépôt. Au final, 48 plaques 384 peuvent être déposées sur une membrane. Pour la banque de clones BAC, il nous faut donc 9 membranes (la dernière ne contient que 4 champs).

Au cours de la duplication, les plaques 98, 99, 107 et 114 ont été perdues et n'ont donc pas pu être déposées sur les membranes. La banque de clones BAC ne contient plus que 408 plaques 384.

Comment lire les résultats d'une hybridation sur ces membranes ?

Prenons l'exemple d'un clone BAC positif sur la membrane F dans le champ 4 dans le puits D15 avec un pattern de type 3. Le pattern de type 3 dans le champ 4 correspond à la plaque 27. La plaque 27 sur la membrane F correspond à la plaque 267 de la banque de clones BAC. Le clone BAC positif est donc le clone BAC contenu dans le puits D15 de la plaque 267 noté $267D15$.

N°plaq	A	B	C	D	E	F	G	H	I
1	1	49	97	145	193	241	289	337	385
2	2	50	–	146	194	242	290	338	386
3	3	51	–	147	195	243	291	339	387
4	4	52	100	148	196	244	292	340	388
5	5	53	101	149	197	245	293	341	389
6	6	54	102	150	198	246	294	342	390
7	7	55	103	151	199	247	295	343	391
8	8	56	104	152	200	248	296	344	392
9	9	57	105	153	201	249	297	345	393
10	10	58	106	154	202	250	298	346	394
11	11	59	–	155	203	251	299	347	395
12	12	60	108	156	204	252	300	348	396
13	13	61	109	157	205	253	301	349	397
14	14	62	110	158	206	254	302	350	398
15	15	63	111	159	207	255	303	351	399
16	16	64	112	160	208	256	304	352	400
17	17	65	113	161	209	257	305	353	401
18	18	66	–	162	210	258	306	354	402
19	19	67	115	163	211	259	307	355	403
20	20	68	116	164	212	260	308	356	404
21	21	69	117	165	213	261	309	357	405
22	22	70	118	166	214	262	310	358	406
23	23	71	119	167	215	263	311	359	407
24	24	72	120	168	216	264	312	360	408
25	25	73	121	169	217	265	313	361	409
26	26	74	122	170	218	266	314	362	410
27	27	75	123	171	219	267	315	363	411
28	28	76	124	172	220	268	316	364	412
29	29	77	125	173	221	269	317	365	34
30	30	78	126	174	222	270	318	366	
31	31	79	127	175	223	271	319	367	
32	32	80	128	176	224	272	320	368	
33	33	81	129	177	225	273	321	369	
34	–	82	130	178	226	274	322	370	
35	35	83	131	179	227	275	323	371	
36	36	84	132	180	228	276	324	372	
37	37	85	133	181	229	277	325	373	
38	38	86	134	182	230	278	326	374	
39	39	87	135	183	231	279	327	375	
40	40	88	136	184	232	280	328	376	
41	41	89	137	185	233	281	329	377	
42	42	90	138	186	234	282	330	378	
43	43	91	139	187	235	283	331	379	
44	44	92	140	188	236	284	332	380	
45	45	93	141	189	237	285	333	381	
46	46	94	142	190	238	286	334	382	
47	47	95	143	191	239	287	335	383	
48	48	96	144	192	240	288	336	384	

TAB. E.1 – Correspondence des plaques déposées sur les membranes.

Annexe F

Résultats des Hybridations des sondes Overgo

Un total de 161 sondes Overgo ont été hybridées en pool de 48, 96 ou 149 sur les 9 membranes qui composent la banque de clones BAC. Les résultats de cette hybridation a permis de récupérer 503 clones BAC positifs qui ont été repiqués et déposés sur de nouvelles membranes. Ces membranes ont servies à identifer les clones BAC spécifiques de chaque sonde. Les sondes qui se sont hybridées avec au moins un clones BAC sont présentées dans les Tableaux ci-dessous (107 sondes Overgo). Cinquante quatre sondes ne sont hybridées avec aucun clone BAC (1, 5, 12, 19, 28, 40, 46, 48, 49, 53, 55, 59, 60, 61, 62, 63, 65, 73, 76, 80, 87, 88, 89, 93, 98, 99, 105, 107, 109, 111, 112, 113, 117, 118, 121, 123, 124, 126, 127, 129, 132, 133, 137, 138, 144, 145, 146, 147, 148, 149, 151, r4, r6, r7 et r10).

Sondes	2	3	4	6	7	8	9	10	11	13	14	15
clones	5E05	5E05 229H19 44P04 3P10 10N08	23P24 89D09 125B04 191M17 229O24 262F03 281H05 302A18 305P05	154J20 240D09 290G10 397O12	93B09 267D15 365D11	133K17 344L10	171F08 134B20 151N12 159E22 171B09 183L02	10N08 44P04	44P04 80K11 211G11 279O18	88J19	10N08 210M10	291P07 298P23 88G06

Sondes	16	17	18	20	21	22	23	24	25	26	27	30
clones	37E12	103O23 272K13 269E13 330E09 406L11	74G01 390A06 363J12	276D06	115C24 233A22 368D03	12F04	412D10	144G21	224N18 108J02	287G02 338P24 381N10 118N12 131P03	274F11	22G23 46G12 141A19 288D18 289D05 357A06

Sondes	32	33	34	35	36	37	38	39	41	42	43	44
clones	34P13 45O12 192E14 200B24 295H23	106H05	7I02	83G22 104N14	161I19 159E22 171B09	17N23 24M14 58P20 73A23 152O22 222C13 284L01 291K21 298F04 299O18 339H19 341C16 376G03	191F17	11G08 53K18 188I19 195G21 196F17 197G05 204M06 219J13 235A08 236D09 246C12 257B15 289L01 288I02 340M22 358F17 364K05 374H20 374K18	63F17 330N24	145J06 278F12	222M17 403C16 309B21	294G04 126F02

Sondes	45	47	50	51	52	54	56	57	58	64	66	67
clones	67D03 211B16	379K10	166B12 354J05	34F18 105B08 157I16	34F18 194F03	312C13	227G18 268F09 279O18 341F23	275F14 335O10 119N17	289D05	265B15 399E03	151N12 151N13 171B09 183L02 386I07 159E22	172M01 197A21

Sondes	68	69	70	71	72	74	75	77	78	79	81	82
clones	144H18 158B23 159E22 171B09 194K11 194M01 201G23 211J08 219B08 298L13 304P21 316J02 316G19 378N09 379I16	2G18 8F16 137C18 138J01 140J02 289M05 399E03	163M02	354M21	97E08 141L19 183I23	326O02	16D13 63G09 410E08	71K06 111K07 157D04 171B09	252G17	276C10 73C20	234K15 29D15 208G04	198E11 205K17

Sondes	83	84	85	86	90	91	92	94	95	96	97	100
clones	164E15	151N12	364P01	287O17	165O02	166P07 50C04	7F22 226L18 284L01 325P22 359B08 374M02 395M03	123P13 176H07 271A23 281N02	123P13 176H07 271A23 281N02	122O11 380O05	34E04 44F21 294M14 295F20 296E09 296F23 299N03 312B08 332I13 339E18 351M14 354H06 136013 159E22 170M22 194K11	269K16 265L16 354J15

Sondes	101	102	103	104	106	108	110	114	115	116	119	120
clones	76L18 269E13 296A06 374G22	23F09 135M03	373D09	385H05	246A07 353L07 394B04	367C18	208J10	94I01 131B05 146H04 368H04	125P03 393I01	206K01	50F11 53E15 92G13	104K04 136E16 255E04 297B18 294G04 339K19 374M02

Sondes	122	125	128	130	131	134	135	136	139	140	141	142
clones	312C01	329N24	14M05	85F23	135G04	38H19	270C08	37I01	190G04	190F14	115I13	58C13
	325C23		72K04	126M12		275G12	367K23	93B09	39J21			95K05
	377F13			126A14		410K18		127C07				210B05
	153N21			139P17		182D02		246P18				287C21
	383A07			145C01		375D07		267D15				393H11
				168K15				292N23				410M04
				198E11				331G09				
				205K17				365D11				
				275I01				369M06				
				316F17								
				333G21								
				348L01								
				348M15								
				353L07								

Sondes	143	150
clones	10L19	12K11
		100C03
		289L01

Sondes	r1	r2	r3	r5	r8	r9	r11	r12
clones	96M06	24K01	191G06	168M23	76C19	203I18	16P05	172M05
			204H17	194I16	145L18	20E25	31M11	387J06
					145J23	150A14	47A17	400H01
					149P10		53P04	
					149O14		184M14	
					169B03		198P17	
					184K17		224H24	
					195I21		230H07	
					245C21		393L17	
					257H06		396M06	
					287O12			
					288C23			
					288B23			
					309H13			
					328C12			
					389I09			
					395K06			
					396I02			
					409J13			
					410H21			

Annexe G

Carte Génétique du Tournesol

Carte génétique du croisement PAC2xRHA266 de Gentzbittel *et al.* (1999).

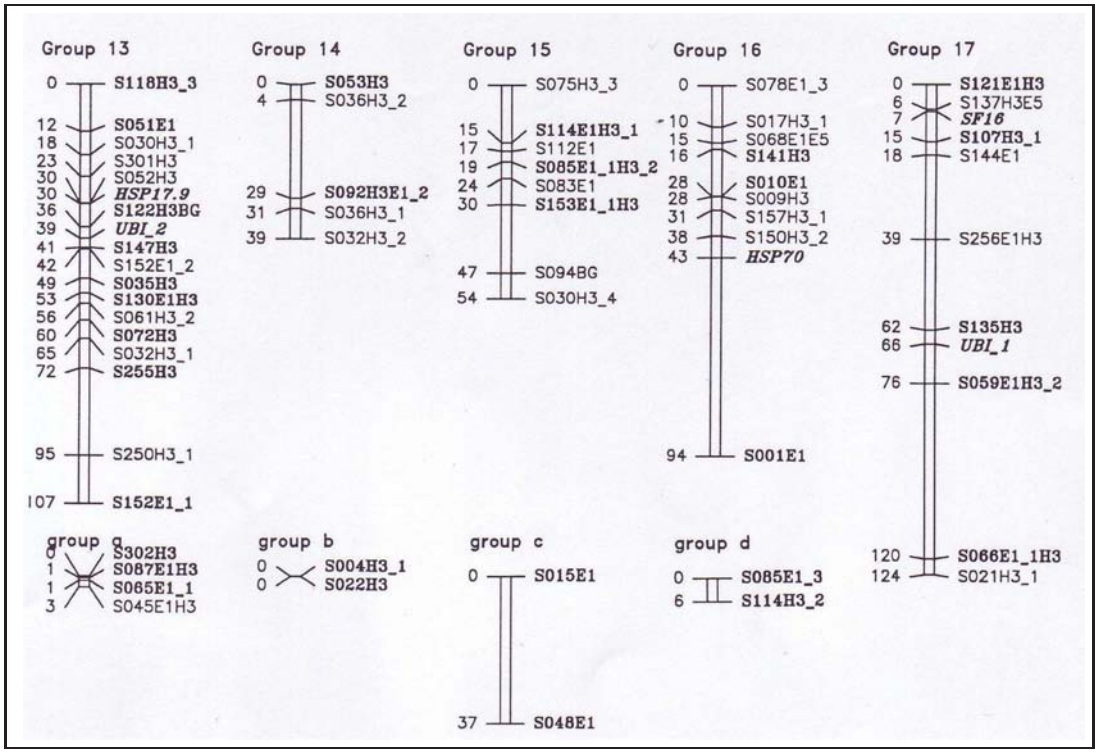


FIG. G.1 – Carte génétique du croisement PAC2xRHA266 d'après Gentzbittel *et al* (1999).

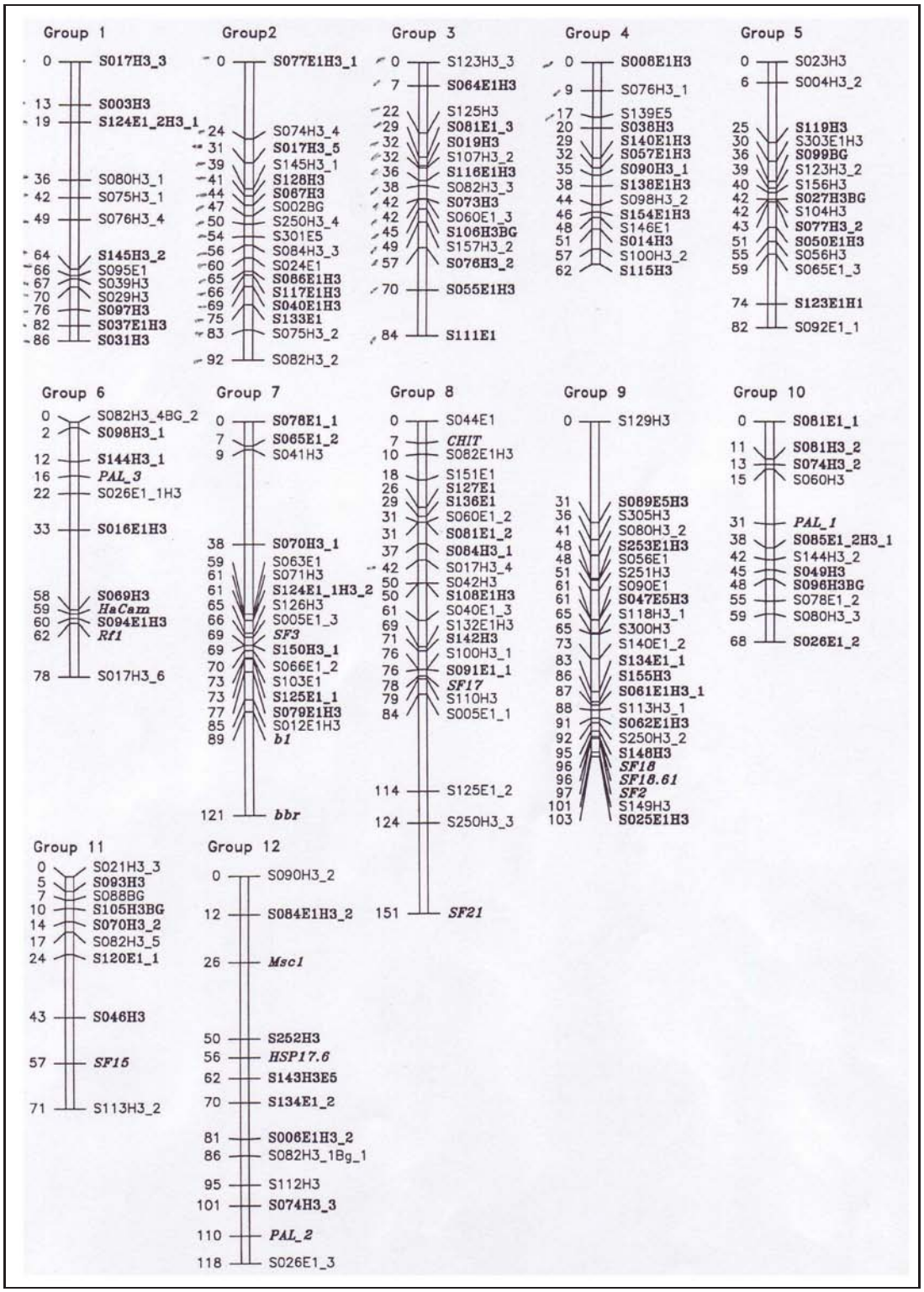


FIG. G.2 – Carte génétique du croisement PAC2xRHA266 d'après Gentzbittel *et al* (1999).

Annexe H

Migration sur Gel d'Agarose 4%

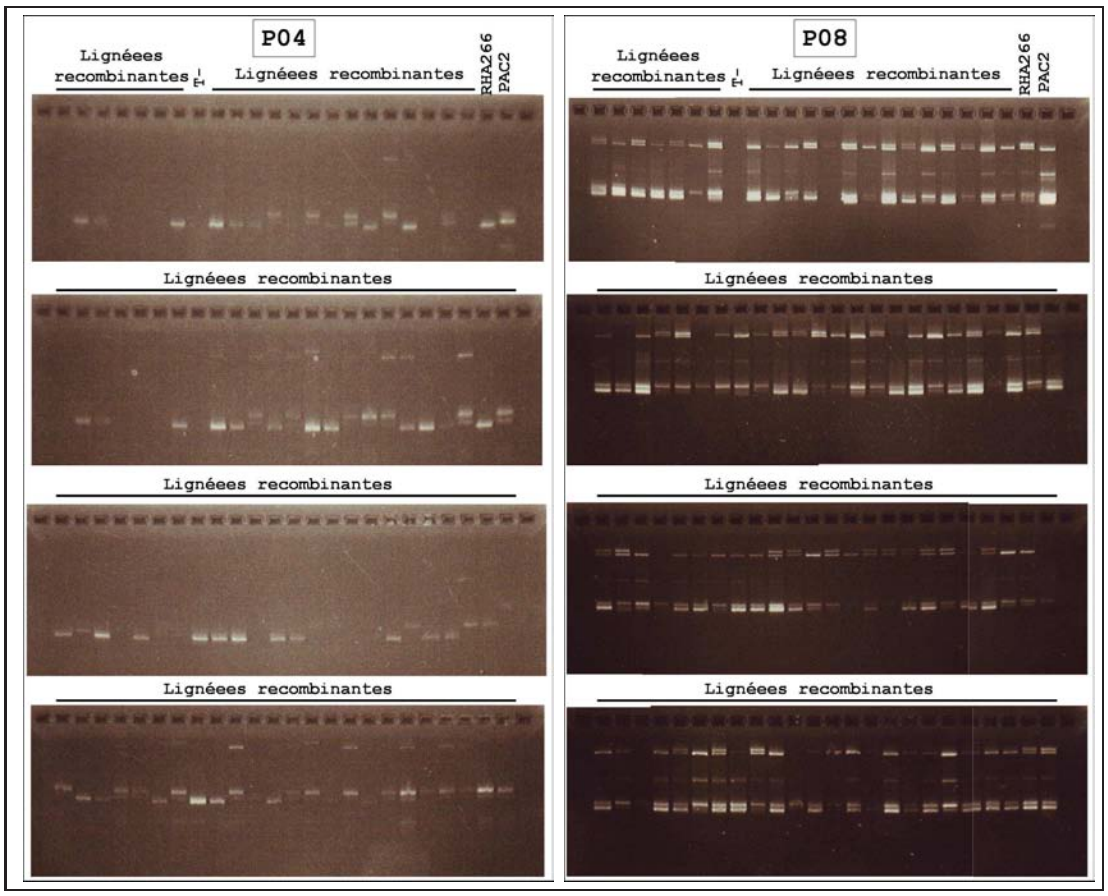


FIG. H.1 – Migration des produits d'amplification des couples d'amorces 4 et 8 à partir des ADN des lignées recombinantes.

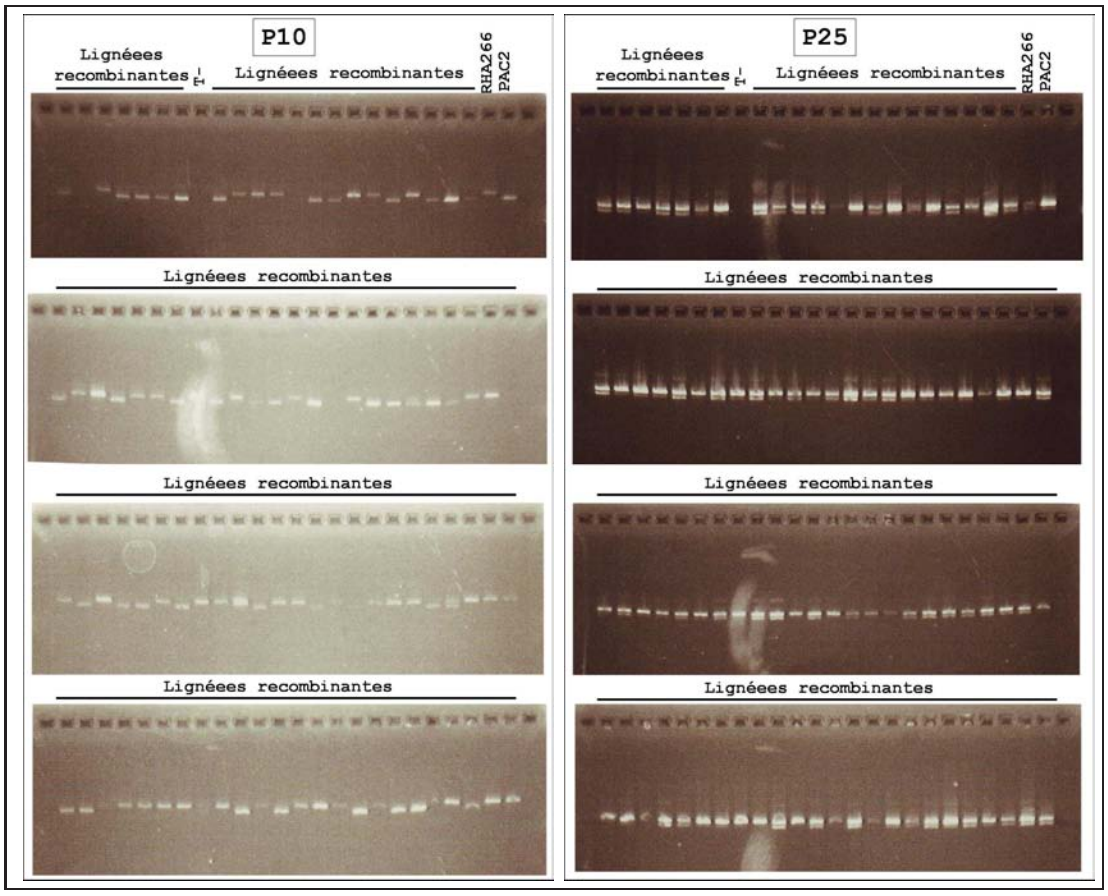


FIG. H.2 – Migration des produits d'amplification des couples d'amorces 10 et 25 à partir des ADN des lignées recombinantes.

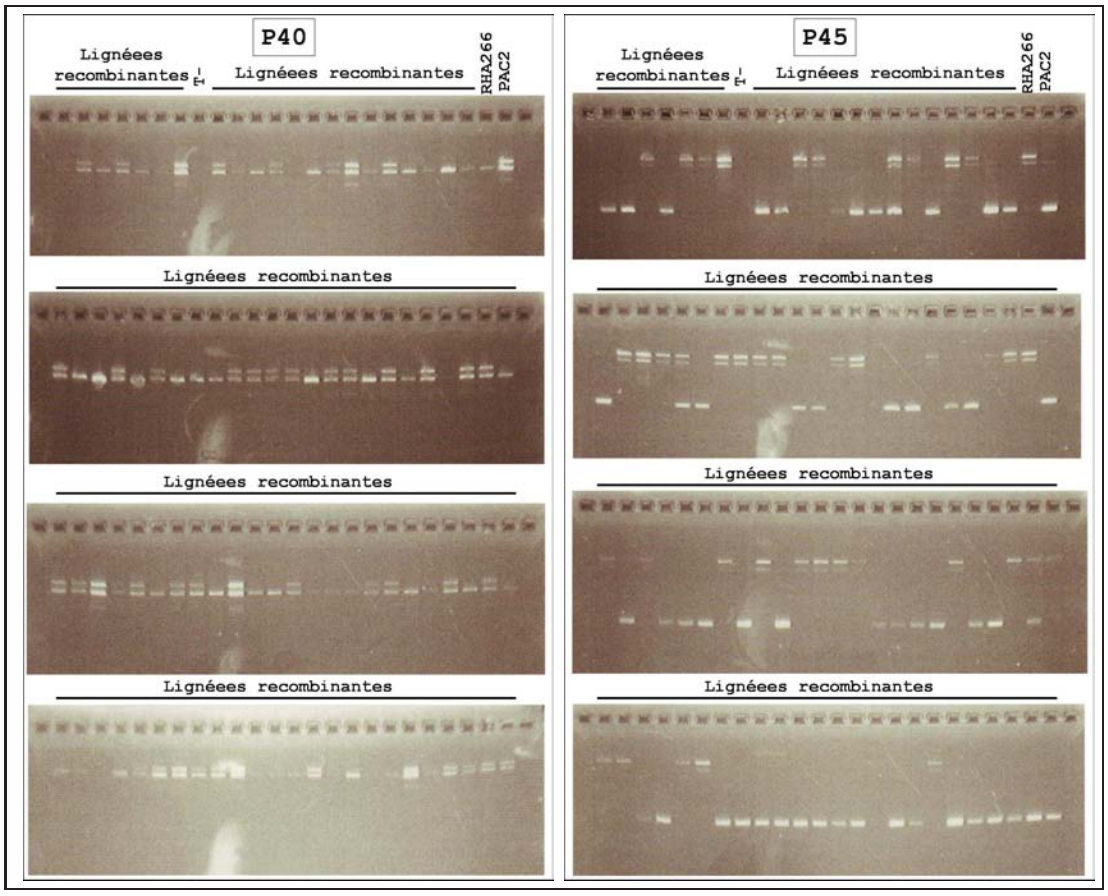


FIG. H.3 – Migration des produits d'amplification des couples d'amorces 40 et 45 à partir des ADN des lignées recombinantes.

Annexe I

Résultats de l'analyse de MapMaker

```
**********************************************************************
* Output from:                                Thu Jun 24 16:51:43 2005 *
*                                                                      *
*                            MAPMAKER/EXP                              *
*                           (version 3.0b)                             *
*                                                                      *
**********************************************************************

data from 'TEST.RAW' are loaded
  RI (selfing) data  (82 individuals, 418 loci)
'photo' is on: file is 'THESE.OUT'

3> sequence all
sequence #1= all

4> group 4
Linkage Groups at min LOD 4.00, max Distance 50.0

group1= 1 17 25 46 133 144 168 193 199 216 226 264 274 284 290 292 307 342 404 405 406
-------
group2= 2 3 32 35 70 93 124 143 194 196 225 247 315 347 372 373 388 418
-------
group3= 4 30 31 51 57 61 75 97 105 127 138 171 198 203 211 233 251 254 272 275 291 305 320 326 334 339 340 383 389 413
-------
group4= 5 201
-------
group5= 6 26 81 85 106 160 172 218 249 267 299 304 341 344 355 362
-------
group6= 8 9 16 20 23 130 162 175 184 205 242 256 298 332 370 393
-------
group7= 10 80
-------
group8= 11 58 109 159 177 325 363
-------
group9= 12 22 27 28 44 56 88 121 132 136 146 153 178 186 232 296 302 396 401 402
-------
group10= 13 89 338 349 374 399 410
-------
group11= 14 15 38 48 52 55 87 92 115 122 128 134 137 140 141 157 176 183 189 190 191 192 202 208 209 210 228 236 237 238 239 244
245 253 263 265 266 276 280 285 308 313 321 350 351 352 364 366 379 380 382 392 394
-------
group12= 18 45 65 72 120 131 142 148 169 181 182 221 227 246 260 270 282 283 297 384 391 395 403 411
-------
group13= 19 42 63 86 111 116 125 152 179 200 288 310 311 322 333 378
-------
group14= 21 49 73 79 119 161 170 206 214 215 217 222 224 229 255 301 309 354 356
361
-------
group15= 24 155 188 357 400
-------
group16= 29 33 76 77 78 100 107 123 163 197 220 248 258 316 324 345 346 358 359 365 367 387 407 415
-------
group17= 36 62 68 94 139 173 219 261 268 281 287 303 317
-------
group18= 37 110 166 417
```

```
-------
group19= 39 41 47 59 101 114 151 164 207 230 234 286 300 306 319 336 348 375 385 386 397 398
-------
group20= 43 60 69 99 102 103 108 117 165 167 212 273 278 293 312 327 328 329 330 360 390 408 409
-------
group21= 50 126 195 250
-------
group22= 53 235 377
-------
group23= 54 71 82 83 90 104 112 113 118 180 185 294 337 369 376
-------
group24= 67 91 149 150 156 174 371
-------
group25= 74 231
-------
group26= 84 289 318
-------
group27= 98 154 331
-------
group28= 129 135 252 279 295
-------
group29= 204 262 353
-------
group30= 213 223
-------
group31= 271 277
-------
group32= 335 368
-------
group33= 343 381
-------
unlinked= 7 34 40 64 66 95 96 145 147 158 187 240 241 243 257 259 269 314 323 412 414 416

5> list loci
                          Error          Linkage   Haplotype
 Num  Name       Genotypes  Prob   Chrom    Group     Group     Class    New?
   1  E32M49_1          68  1.00\%  -       group1    -         -        -
   2  E32M49_2          67  1.00\%  -       group2    -         -        -
   3  E32M49_5          77  1.00\%  -       group2    -         -        -
   4  E32M49_7          78  1.00\%  -       group3    -         -        -
   5  E32M49_9          78  1.00\%  -       group4    -         -        -
   6  E32M49_1          62  1.00\%  -       group5    -         -        -
   7  E32M49_1          60  1.00\%  -       unlinked  -         -        -
   8  E32M49_1          76  1.00\%  -       group6    -         -        -
   9  E32M49_2          79  1.00\%  -       group6    -         -        -
  10  E32M49_2          79  1.00\%  -       group7    -         -        -
  11  E32M49_2          70  1.00\%  -       group8    -         -        -
  12  E32M49_2          78  1.00\%  -       group9    -         -        -
  13  E32M49_2          78  1.00\%  -       group10   -         -        -
  14  E32M49_3          63  1.00\%  -       group11   -         -        -
  15  E32M49_3          76  1.00\%  -       group11   -         -        -
  16  E32M49_3          76  1.00\%  -       group6    -         -        -
  17  E33M62_6          76  1.00\%  -       group1    -         -        -
  18  E33M62_9          78  1.00\%  -       group12   -         -        -
  19  E33M62_1          76  1.00\%  -       group13   -         -        -
  20  E40M59_3          78  1.00\%  -       group6    -         -        -
  21  E40M59_7          77  1.00\%  -       group14   -         -        -
  22  E40M59_8          67  1.00\%  -       group9    -         -        -
  23  E40M59_9          78  1.00\%  -       group6    -         -        -
  24  E40M59_1          76  1.00\%  -       group15   -         -        -
  25  E40M59_1          78  1.00\%  -       group1    -         -        -
  26  E38M50_2          72  1.00\%  -       group5    -         -        -
  27  E38M50_4          74  1.00\%  -       group9    -         -        -
  28  E38M50_6          71  1.00\%  -       group9    -         -        -
  29  E38M50_7          70  1.00\%  -       group16   -         -        -
  30  E38M50_1          77  1.00\%  -       group3    -         -        -
  31  E38M50_1          72  1.00\%  -       group3    -         -        -
  32  E38M50_1          77  1.00\%  -       group2    -         -        -
  33  E38M50_2          77  1.00\%  -       group16   -         -        -
  34  E38M50_2          77  1.00\%  -       unlinked  -         -        -
  35  E38M50_2          66  1.00\%  -       group2    -         -        -
  36  E38M50_2          73  1.00\%  -       group17   -         -        -
  37  E37M47_3          76  1.00\%  -       group18   -         -        -
  38  E37M47_4          74  1.00\%  -       group11   -         -        -
  39  E37M47_5          77  1.00\%  -       group19   -         -        -
  40  E37M47_7          76  1.00\%  -       unlinked  -         -        -
  41  E37M47_1          75  1.00\%  -       group19   -         -        -
  42  E37M47_1          77  1.00\%  -       group13   -         -        -
  43  E37M47_1          73  1.00\%  -       group20   -         -        -
  44  E37M47_2          65  1.00\%  -       group9    -         -        -
  45  E37M47_2          66  1.00\%  -       group12   -         -        -
  46  E37M47_2          66  1.00\%  -       group1    -         -        -
  47  E37M47_2          77  1.00\%  -       group19   -         -        -
  48  E41M59_1          77  1.00\%  -       group11   -         -        -
```

49	E41M59_3	77	1.00\%	-	group14	-	-	-
50	E41M59_4	74	1.00\%	-	group21	-	-	-
51	E41M59_5	77	1.00\%	-	group3	-	-	-
52	E41M59_6	76	1.00\%	-	group11	-	-	-
53	E41M59_7	77	1.00\%	-	group22	-	-	-
54	E41M59_1	66	1.00\%	-	group23	-	-	-
55	E37M49_5	79	1.00\%	-	group11	-	-	-
56	E37M49_6	79	1.00\%	-	group9	-	-	-
57	E37M49_7	79	1.00\%	-	group3	-	-	-
58	E37M49_9	74	1.00\%	-	group8	-	-	-
59	E32M47_1	79	1.00\%	-	group19	-	-	-
60	E32M47_2	79	1.00\%	-	group20	-	-	-
61	E32M47_8	80	1.00\%	-	group3	-	-	-
62	E32M47_9	80	1.00\%	-	group17	-	-	-
63	E32M47_1	80	1.00\%	-	group13	-	-	-
64	E32M47_1	78	1.00\%	-	unlinked	-	-	-
65	E36M59_3	80	1.00\%	-	group12	-	-	-
66	E36M59_4	80	1.00\%	-	unlinked	-	-	-
67	E36M59_5	76	1.00\%	-	group24	-	-	-
68	E36M59_7	69	1.00\%	-	group17	-	-	-
69	E36M59_1	80	1.00\%	-	group20	-	-	-
70	E41M62_4	72	1.00\%	-	group2	-	-	-
71	E41M62_7	73	1.00\%	-	group23	-	-	-
72	E41M62_9	76	1.00\%	-	group12	-	-	-
73	E41M62_1	76	1.00\%	-	group14	-	-	-
74	E41M62_1	76	1.00\%	-	group25	-	-	-
75	E41M62_1	76	1.00\%	-	group3	-	-	-
76	E41M62_1	75	1.00\%	-	group16	-	-	-
77	E41M62_2	76	1.00\%	-	group16	-	-	-
78	E41M62_2	68	1.00\%	-	group16	-	-	-
79	E41M62_2	76	1.00\%	-	group14	-	-	-
80	E41M62_2	74	1.00\%	-	group7	-	-	-
81	E41M62_2	77	1.00\%	-	group5	-	-	-
82	E41M62_3	77	1.00\%	-	group23	-	-	-
83	E35M60_1	78	1.00\%	-	group23	-	-	-
84	E35M60_4	76	1.00\%	-	group26	-	-	-
85	E35M60_1	77	1.00\%	-	group5	-	-	-
86	E35M60_1	67	1.00\%	-	group13	-	-	-
87	E35M60_1	78	1.00\%	-	group11	-	-	-
88	E35M60_1	78	1.00\%	-	group9	-	-	-
89	E35M60_1	78	1.00\%	-	group10	-	-	-
90	E35M60_1	66	1.00\%	-	group23	-	-	-
91	E35M60_1	76	1.00\%	-	group24	-	-	-
92	E35M60_2	67	1.00\%	-	group11	-	-	-
93	E35M60_2	77	1.00\%	-	group2	-	-	-
94	E40M47_3	78	1.00\%	-	group17	-	-	-
95	E40M47_4	77	1.00\%	-	unlinked	-	-	-
96	E40M47_9	74	1.00\%	-	unlinked	-	-	-
97	E40M47_1	78	1.00\%	-	group3	-	-	-
98	E40M47_1	78	1.00\%	-	group27	-	-	-
99	E40M47_1	77	1.00\%	-	group20	-	-	-
100	E40M47_2	72	1.00\%	-	group16	-	-	-
101	E37M61_1	77	1.00\%	-	group19	-	-	-
102	E37M61_2	78	1.00\%	-	group20	-	-	-
103	E37M61_7	78	1.00\%	-	group20	-	-	-
104	E37M61_1	75	1.00\%	-	group23	-	-	-
105	E37M61_1	78	1.00\%	-	group3	-	-	-
106	E33M48_1	76	1.00\%	-	group5	-	-	-
107	E33M48_3	82	1.00\%	-	group16	-	-	-
108	E33M48_5	82	1.00\%	-	group20	-	-	-
109	E33M48_6	80	1.00\%	-	group8	-	-	-
110	E33M48_1	77	1.00\%	-	group18	-	-	-
111	E33M48_1	81	1.00\%	-	group13	-	-	-
112	E33M48_2	81	1.00\%	-	group23	-	-	-
113	E33M48_2	81	1.00\%	-	group23	-	-	-
114	E33M48_2	80	1.00\%	-	group19	-	-	-
115	E40M62_2	77	1.00\%	-	group11	-	-	-
116	E40M62_4	78	1.00\%	-	group13	-	-	-
117	E40M62_7	80	1.00\%	-	group20	-	-	-
118	E40M62_8	80	1.00\%	-	group23	-	-	-
119	E40M62_1	80	1.00\%	-	group14	-	-	-
120	E40M62_1	79	1.00\%	-	group12	-	-	-
121	E40M62_1	68	1.00\%	-	group9	-	-	-
122	E40M62_1	80	1.00\%	-	group11	-	-	-
123	E40M62_1	80	1.00\%	-	group16	-	-	-
124	E40M62_1	79	1.00\%	-	group2	-	-	-
125	E40M62_1	79	1.00\%	-	group13	-	-	-
126	E40M62_2	67	1.00\%	-	group21	-	-	-
127	E40M62_2	80	1.00\%	-	group3	-	-	-
128	E40M62_2	80	1.00\%	-	group11	-	-	-
129	E40M62_2	65	1.00\%	-	group28	-	-	-
130	E40M50_5	70	1.00\%	-	group6	-	-	-
131	E40M50_6	71	1.00\%	-	group12	-	-	-
132	E40M50_9	71	1.00\%	-	group9	-	-	-

133	E40M50_1	59	1.00\%	-	group1	-	-	-
134	E40M50_1	59	1.00\%	-	group11	-	-	-
135	E40M50_1	71	1.00\%	-	group28	-	-	-
136	E40M50_1	70	1.00\%	-	group9	-	-	-
137	E40M50_1	71	1.00\%	-	group11	-	-	-
138	E40M50_1	71	1.00\%	-	group3	-	-	-
139	E41M50_2	76	1.00\%	-	group17	-	-	-
140	E41M50_4	76	1.00\%	-	group11	-	-	-
141	E41M50_7	75	1.00\%	-	group11	-	-	-
142	E41M50_9	76	1.00\%	-	group12	-	-	-
143	E41M50_1	60	1.00\%	-	group2	-	-	-
144	E41M50_1	61	1.00\%	-	group1	-	-	-
145	E32M61_8	79	1.00\%	-	unlinked	-	-	-
146	E32M61_9	79	1.00\%	-	group9	-	-	-
147	E32M61_1	80	1.00\%	-	unlinked	-	-	-
148	E32M61_1	80	1.00\%	-	group12	-	-	-
149	E32M61_1	80	1.00\%	-	group24	-	-	-
150	E32M61_1	73	1.00\%	-	group24	-	-	-
151	E38M48_2	77	1.00\%	-	group19	-	-	-
152	E38M48_3	77	1.00\%	-	group13	-	-	-
153	E38M48_6	77	1.00\%	-	group9	-	-	-
154	E38M48_7	77	1.00\%	-	group27	-	-	-
155	E38M48_1	73	1.00\%	-	group15	-	-	-
156	E38M48_1	70	1.00\%	-	group24	-	-	-
157	E33M60_2	71	1.00\%	-	group11	-	-	-
158	E33M60_4	70	1.00\%	-	unlinked	-	-	-
159	E33M60_6	71	1.00\%	-	group8	-	-	-
160	E33M60_7	71	1.00\%	-	group5	-	-	-
161	E33M60_8	71	1.00\%	-	group14	-	-	-
162	E41M48_2	56	1.00\%	-	group6	-	-	-
163	E41M48_3	56	1.00\%	-	group16	-	-	-
164	E41M48_4	57	1.00\%	-	group19	-	-	-
165	E41M48_7	61	1.00\%	-	group20	-	-	-
166	E41M48_8	59	1.00\%	-	group18	-	-	-
167	E41M48_9	59	1.00\%	-	group20	-	-	-
168	E41M48_1	63	1.00\%	-	group1	-	-	-
169	E41M48_1	60	1.00\%	-	group12	-	-	-
170	E38M60_3	62	1.00\%	-	group14	-	-	-
171	E38M60_6	63	1.00\%	-	group3	-	-	-
172	E38M60_8	67	1.00\%	-	group5	-	-	-
173	E38M60_1	58	1.00\%	-	group17	-	-	-
174	E38M60_1	60	1.00\%	-	group24	-	-	-
175	E38M60_1	65	1.00\%	-	group6	-	-	-
176	E33M50_1	71	1.00\%	-	group11	-	-	-
177	E33M50_5	67	1.00\%	-	group8	-	-	-
178	E35M61_2	65	1.00\%	-	group9	-	-	-
179	E35M61_3	65	1.00\%	-	group13	-	-	-
180	E35M61_7	67	1.00\%	-	group23	-	-	-
181	E35M61_8	67	1.00\%	-	group12	-	-	-
182	E35M61_1	65	1.00\%	-	group12	-	-	-
183	E35M61_1	62	1.00\%	-	group11	-	-	-
184	E35M49_5	56	1.00\%	-	group6	-	-	-
185	E35M49_6	57	1.00\%	-	group23	-	-	-
186	E35M49_7	57	1.00\%	-	group9	-	-	-
187	E35M49_8	55	1.00\%	-	unlinked	-	-	-
188	E35M49_1	49	1.00\%	-	group15	-	-	-
189	E35M62_3	69	1.00\%	-	group11	-	-	-
190	E35M62_4	68	1.00\%	-	group11	-	-	-
191	E35M62_7	70	1.00\%	-	group11	-	-	-
192	E35M62_9	67	1.00\%	-	group11	-	-	-
193	E35M62_1	70	1.00\%	-	group1	-	-	-
194	E35M48_3	72	1.00\%	-	group2	-	-	-
195	E35M48_6	66	1.00\%	-	group21	-	-	-
196	E35M48_1	71	1.00\%	-	group2	-	-	-
197	E35M48_1	68	1.00\%	-	group16	-	-	-
198	E35M48_1	69	1.00\%	-	group3	-	-	-
199	E32M49_3	71	1.00\%	-	group1	-	-	-
200	E32M49_8	76	1.00\%	-	group13	-	-	-
201	E32M49_1	78	1.00\%	-	group4	-	-	-
202	E32M49_1	77	1.00\%	-	group11	-	-	-
203	E32M49_2	76	1.00\%	-	group3	-	-	-
204	E32M49_2	73	1.00\%	-	group29	-	-	-
205	E32M49_2	78	1.00\%	-	group6	-	-	-
206	E32M49_2	77	1.00\%	-	group14	-	-	-
207	E32M49_2	77	1.00\%	-	group19	-	-	-
208	E32M49_3	77	1.00\%	-	group11	-	-	-
209	E32M49_3	77	1.00\%	-	group11	-	-	-
210	E33M62_7	77	1.00\%	-	group11	-	-	-
211	E33M62_8	79	1.00\%	-	group3	-	-	-
212	E33M62_1	78	1.00\%	-	group20	-	-	-
213	E40M59_4	68	1.00\%	-	group30	-	-	-
214	E40M59_5	79	1.00\%	-	group14	-	-	-
215	E40M59_6	79	1.00\%	-	group14	-	-	-
216	E40M59_1	78	1.00\%	-	group1	-	-	-

217	E40M59_1	77	1.00\%	-	group14	-	-	-
218	E38M50_1	71	1.00\%	-	group5	-	-	-
219	E38M50_3	72	1.00\%	-	group17	-	-	-
220	E38M50_5	75	1.00\%	-	group16	-	-	-
221	E38M50_8	75	1.00\%	-	group12	-	-	-
222	E38M50_9	75	1.00\%	-	group14	-	-	-
223	E38M50_1	67	1.00\%	-	group30	-	-	-
224	E38M50_1	77	1.00\%	-	group14	-	-	-
225	E38M50_1	78	1.00\%	-	group2	-	-	-
226	E38M50_1	77	1.00\%	-	group1	-	-	-
227	E38M50_2	76	1.00\%	-	group12	-	-	-
228	E37M47_2	75	1.00\%	-	group11	-	-	-
229	E37M47_8	76	1.00\%	-	group14	-	-	-
230	E37M47_9	75	1.00\%	-	group19	-	-	-
231	E37M47_1	76	1.00\%	-	group25	-	-	-
232	E37M47_1	77	1.00\%	-	group9	-	-	-
233	E37M47_1	75	1.00\%	-	group3	-	-	-
234	E41M59_2	74	1.00\%	-	group19	-	-	-
235	E41M59_8	77	1.00\%	-	group22	-	-	-
236	E41M59_9	77	1.00\%	-	group11	-	-	-
237	E41M59_1	77	1.00\%	-	group11	-	-	-
238	E37M49_3	77	1.00\%	-	group11	-	-	-
239	E37M49_4	79	1.00\%	-	group11	-	-	-
240	E32M47_7	80	1.00\%	-	unlinked	-	-	-
241	E32M47_1	80	1.00\%	-	unlinked	-	-	-
242	E32M47_1	79	1.00\%	-	group6	-	-	-
243	E36M59_1	77	1.00\%	-	unlinked	-	-	-
244	E36M59_2	69	1.00\%	-	group11	-	-	-
245	E36M59_6	80	1.00\%	-	group11	-	-	-
246	E36M59_8	80	1.00\%	-	group12	-	-	-
247	E36M59_9	80	1.00\%	-	group2	-	-	-
248	E36M59_1	80	1.00\%	-	group16	-	-	-
249	E36M59_1	80	1.00\%	-	group5	-	-	-
250	E36M59_1	78	1.00\%	-	group21	-	-	-
251	E36M59_1	80	1.00\%	-	group3	-	-	-
252	E36M59_1	80	1.00\%	-	group28	-	-	-
253	E41M62_1	63	1.00\%	-	group11	-	-	-
254	E41M62_2	63	1.00\%	-	group3	-	-	-
255	E41M62_6	76	1.00\%	-	group14	-	-	-
256	E41M62_1	76	1.00\%	-	group6	-	-	-
257	E41M62_1	76	1.00\%	-	unlinked	-	-	-
258	E41M62_1	75	1.00\%	-	group16	-	-	-
259	E41M62_2	73	1.00\%	-	unlinked	-	-	-
260	E41M62_2	74	1.00\%	-	group12	-	-	-
261	E41M62_2	77	1.00\%	-	group17	-	-	-
262	E35M60_3	77	1.00\%	-	group29	-	-	-
263	E35M60_5	78	1.00\%	-	group11	-	-	-
264	E35M60_6	67	1.00\%	-	group1	-	-	-
265	E35M60_7	76	1.00\%	-	group11	-	-	-
266	E35M60_1	78	1.00\%	-	group11	-	-	-
267	E35M60_2	77	1.00\%	-	group5	-	-	-
268	E35M60_2	75	1.00\%	-	group17	-	-	-
269	E40M47_1	77	1.00\%	-	unlinked	-	-	-
270	E40M47_2	78	1.00\%	-	group12	-	-	-
271	E40M47_5	66	1.00\%	-	group31	-	-	-
272	E40M47_7	77	1.00\%	-	group3	-	-	-
273	E40M47_8	78	1.00\%	-	group20	-	-	-
274	E40M47_1	78	1.00\%	-	group1	-	-	-
275	E40M47_1	78	1.00\%	-	group3	-	-	-
276	E40M47_1	78	1.00\%	-	group11	-	-	-
277	E40M47_1	78	1.00\%	-	group31	-	-	-
278	E40M47_1	78	1.00\%	-	group20	-	-	-
279	E40M47_2	72	1.00\%	-	group28	-	-	-
280	E40M47_2	77	1.00\%	-	group11	-	-	-
281	E40M47_2	72	1.00\%	-	group17	-	-	-
282	E37M61_3	78	1.00\%	-	group12	-	-	-
283	E37M61_4	78	1.00\%	-	group12	-	-	-
284	E37M61_5	77	1.00\%	-	group1	-	-	-
285	E37M61_6	77	1.00\%	-	group11	-	-	-
286	E37M61_8	75	1.00\%	-	group19	-	-	-
287	E37M61_9	75	1.00\%	-	group17	-	-	-
288	E33M48_2	82	1.00\%	-	group13	-	-	-
289	E33M48_4	78	1.00\%	-	group26	-	-	-
290	E33M48_8	81	1.00\%	-	group1	-	-	-
291	E33M48_9	81	1.00\%	-	group3	-	-	-
292	E33M48_1	81	1.00\%	-	group1	-	-	-
293	E33M48_1	80	1.00\%	-	group20	-	-	-
294	E33M48_1	81	1.00\%	-	group23	-	-	-
295	E33M48_2	77	1.00\%	-	group28	-	-	-
296	E33M48_2	80	1.00\%	-	group9	-	-	-
297	E33M48_2	81	1.00\%	-	group12	-	-	-
298	E33M48_2	79	1.00\%	-	group6	-	-	-
299	E40M62_1	65	1.00\%	-	group5	-	-	-
300	E40M62_5	68	1.00\%	-	group19	-	-	-

301	E40M62_1	80	1.00\%	-	group14	-	-	-
302	E40M62_1	68	1.00\%	-	group9	-	-	-
303	E40M62_1	66	1.00\%	-	group17	-	-	-
304	E40M62_2	80	1.00\%	-	group5	-	-	-
305	E40M62_2	79	1.00\%	-	group3	-	-	-
306	E40M50_1	71	1.00\%	-	group19	-	-	-
307	E40M50_2	71	1.00\%	-	group1	-	-	-
308	E40M50_8	71	1.00\%	-	group11	-	-	-
309	E40M50_1	70	1.00\%	-	group14	-	-	-
310	E40M50_1	71	1.00\%	-	group13	-	-	-
311	E40M50_1	71	1.00\%	-	group13	-	-	-
312	E40M50_2	71	1.00\%	-	group20	-	-	-
313	E41M50_3	76	1.00\%	-	group11	-	-	-
314	E41M50_5	30	1.00\%	-	unlinked	-	-	-
315	E41M50_6	75	1.00\%	-	group2	-	-	-
316	E41M50_1	76	1.00\%	-	group16	-	-	-
317	E41M50_1	58	1.00\%	-	group17	-	-	-
318	E32M61_2	77	1.00\%	-	group26	-	-	-
319	E32M61_5	80	1.00\%	-	group19	-	-	-
320	E32M61_6	80	1.00\%	-	group3	-	-	-
321	E32M61_7	80	1.00\%	-	group11	-	-	-
322	E32M61_1	79	1.00\%	-	group13	-	-	-
323	E32M61_1	80	1.00\%	-	unlinked	-	-	-
324	E38M48_1	77	1.00\%	-	group16	-	-	-
325	E38M48_4	78	1.00\%	-	group8	-	-	-
326	E38M48_5	77	1.00\%	-	group3	-	-	-
327	E38M48_8	76	1.00\%	-	group20	-	-	-
328	E38M48_9	76	1.00\%	-	group20	-	-	-
329	E38M48_1	71	1.00\%	-	group20	-	-	-
330	E38M48_1	70	1.00\%	-	group20	-	-	-
331	E38M48_1	70	1.00\%	-	group27	-	-	-
332	E33M60_1	67	1.00\%	-	group6	-	-	-
333	E33M60_3	70	1.00\%	-	group13	-	-	-
334	E33M60_5	67	1.00\%	-	group3	-	-	-
335	E41M48_1	54	1.00\%	-	group32	-	-	-
336	E41M48_6	61	1.00\%	-	group19	-	-	-
337	E41M48_1	64	1.00\%	-	group23	-	-	-
338	E38M60_2	63	1.00\%	-	group10	-	-	-
339	E38M60_4	60	1.00\%	-	group3	-	-	-
340	E38M60_5	60	1.00\%	-	group3	-	-	-
341	E38M60_7	56	1.00\%	-	group5	-	-	-
342	E38M60_9	59	1.00\%	-	group1	-	-	-
343	E38M60_1	65	1.00\%	-	group33	-	-	-
344	E33M50_2	68	1.00\%	-	group5	-	-	-
345	E33M50_3	66	1.00\%	-	group16	-	-	-
346	E33M50_4	67	1.00\%	-	group16	-	-	-
347	E33M50_6	66	1.00\%	-	group2	-	-	-
348	E35M61_4	65	1.00\%	-	group19	-	-	-
349	E35M61_5	66	1.00\%	-	group10	-	-	-
350	E35M61_6	67	1.00\%	-	group11	-	-	-
351	E35M61_9	67	1.00\%	-	group11	-	-	-
352	E35M61_1	62	1.00\%	-	group11	-	-	-
353	E35M61_1	63	1.00\%	-	group29	-	-	-
354	E35M49_4	53	1.00\%	-	group14	-	-	-
355	E35M49_9	55	1.00\%	-	group5	-	-	-
356	E35M62_1	68	1.00\%	-	group14	-	-	-
357	E35M62_2	69	1.00\%	-	group15	-	-	-
358	E35M62_5	69	1.00\%	-	group16	-	-	-
359	E35M62_8	69	1.00\%	-	group16	-	-	-
360	E35M62_1	70	1.00\%	-	group20	-	-	-
361	E35M62_1	70	1.00\%	-	group14	-	-	-
362	E35M48_1	69	1.00\%	-	group5	-	-	-
363	E35M48_4	72	1.00\%	-	group8	-	-	-
364	E35M48_5	73	1.00\%	-	group11	-	-	-
365	E35M48_7	73	1.00\%	-	group16	-	-	-
366	E35M48_8	73	1.00\%	-	group11	-	-	-
367	E35M48_9	69	1.00\%	-	group16	-	-	-
368	E35M48_1	67	1.00\%	-	group32	-	-	-
369	E35M48_1	72	1.00\%	-	group23	-	-	-
370	E35M48_1	69	1.00\%	-	group6	-	-	-
371	E35M48_1	71	1.00\%	-	group24	-	-	-
372	ORS5	71	1.00\%	-	group2	-	-	-
373	ORS5fluo	46	1.00\%	-	group2	-	-	-
374	ORS8	70	1.00\%	-	group10	-	-	-
375	ORS31_1	76	1.00\%	-	group19	-	-	-
376	ORS31_2	74	1.00\%	-	group23	-	-	-
377	ORS31_3	72	1.00\%	-	group22	-	-	-
378	ORS53	69	1.00\%	-	group13	-	-	-
379	ORS78	74	1.00\%	-	group11	-	-	-
380	ORS78_k1	61	1.00\%	-	group11	-	-	-
381	iub_6	55	1.00\%	-	group33	-	-	-
382	SSL3	59	1.00\%	-	group11	-	-	-
383	SSL13	63	1.00\%	-	group3	-	-	-
384	SSL20_1	76	1.00\%	-	group12	-	-	-

```
 385  SSL20_2          80  1.00\%  -        group19   -          -        -
 386  SSL22_1          60  1.00\%  -        group19   -          -        -
 387  SSL22_2          72  1.00\%  -        group16   -          -        -
 388  SSL27            57  1.00\%  -        group2    -          -        -
 389  SSL29            73  1.00\%  -        group3    -          -        -
 390  SSL30            73  1.00\%  -        group20   -          -        -
 391  SSL33            73  1.00\%  -        group12   -          -        -
 392  SSL49            71  1.00\%  -        group11   -          -        -
 393  SSL66_1          74  1.00\%  -        group6    -          -        -
 394  SSL66_2          67  1.00\%  -        group11   -          -        -
 395  ORS123_1         61  1.00\%  -        group12   -          -        -
 396  ORS123_2         76  1.00\%  -        group9    -          -        -
 397  ORS126           55  1.00\%  -        group19   -          -        -
 398  ORS128           67  1.00\%  -        group19   -          -        -
 399  ORS121           64  1.00\%  -        group10   -          -        -
 400  SSU39            62  1.00\%  -        group15   -          -        -
 401  SSU41            60  1.00\%  -        group9    -          -        -
 402  SSU100           72  1.00\%  -        group9    -          -        -
 403  SSU123_1         69  1.00\%  -        group12   -          -        -
 404  SSU123_2         68  1.00\%  -        group1    -          -        -
 405  SSU129_1         76  1.00\%  -        group1    -          -        -
 406  SSU129_2         73  1.00\%  -        group1    -          -        -
 407  ORS169           76  1.00\%  -        group16   -          -        -
 408  SSU195           74  1.00\%  -        group20   -          -        -
 409  SSU217           71  1.00\%  -        group20   -          -        -
 410  SSU223           74  1.00\%  -        group10   -          -        -
 411  SSU227           74  1.00\%  -        group12   -          -        -
 412  P8               76  1.00\%  -        unlinked  -          -        -
 413  P10              73  1.00\%  -        group3    -          -        -
 414  P15              46  1.00\%  -        unlinked  -          -        -
 415  P25              82  1.00\%  -        group16   -          -        -
 416  P40              80  1.00\%  -        unlinked  -          -        -
 417  P45              77  1.00\%  -        group18   -          -        -
 418  P4               68  1.00\%  -        group2    -          -        -

6> sequence 124 347 194 372 315 2 93 3 196 143 388 70 32 247 225 35
sequence #3= 124 347 194 372 315 2 93 3 196 143 388 70 32 247 225 35

8> try 418

        418
       -------
      | -4.49 |
124   |       |
      | -8.14 |
347   |       |
      |-12.11 |
194   |       |
      |-12.48 |
372   |       |
      |-18.74 |
315   |       |
      |-17.12 |
2     |       |
      |-15.03 |
93    |       |
      |-13.02 |
3     |       |
      |-10.69 |
196   |       |
      | -8.42 |
143   |       |
      |-10.53 |
388   |       |
      |-11.96 |
70    |       |
      |-10.15 |
32    |       |
      |-12.26 |
247   |       |
      | -2.79 |
225   |       |
      | -9.05 |
35    |       |
      |  0.00 |
      |-------|
INF   | -5.65 |
       -------
BEST  -209.68

9> sequence 124 347 194 372 315 2 93 3 196 143 388 70 32 247 225 35 418
sequence #4= 124 347 194 372 315 2 93 3 196 143 388 70 32 247 225 35 418

10> map
```

```
==========================================================================
Map:
  Markers          Distance
  124  E40M62_1     18.0 cM
  347  E33M50_6      7.5 cM
  194  E35M48_3      8.1 cM
  372  ORS5          1.6 cM
  315  E41M50_6      2.8 cM
    2  E32M49_2      5.2 cM
   93  E35M60_2      9.7 cM
    3  E32M49_5     10.7 cM
  196  E35M48_1      9.8 cM
  143  E41M50_1      4.1 cM
  388  SSL27         6.5 cM
   70  E41M62_4      9.5 cM
   32  E38M50_1      6.3 cM
  247  E36M59_9     14.5 cM
  225  E38M50_1      1.7 cM
   35  E38M50_2     13.4 cM
  418  P4          ----------
                    129.2 cM   17 markers   log-likelihood= -209.68
==========================================================================

11> sequence 61 233 326 320 334 203 198 31 30 51 4 211 251 254 305 127 57 105 291 138 275 272 97 339 171 389 383 340
sequence #6= 61 233 326 320 334 203 198 31 30 51 4 211 251 254 305 127 57 105 291 138 275 272 97 339 171 389 383 340

12> try 413

        413
       -------
      | -6.26 |
61    |       |
      |-10.44 |
233   |       |
      |-15.81 |
326   |       |
      |-13.49 |
320   |       |
      |-10.21 |
334   |       |
      |-10.74 |
203   |       |
      | -9.86 |
198   |       |
      |-11.00 |
31    |       |
      |-14.19 |
30    |       |
      |-13.02 |
51    |       |
      |-14.99 |
4     |       |
      |-14.21 |
211   |       |
      |-12.08 |
251   |       |
      | -8.25 |
254   |       |
      | -3.77 |
305   |       |
      | -9.11 |
127   |       |
      | -3.66 |
57    |       |
      | -3.15 |
105   |       |
      | -3.34 |
291   |       |
      | -5.78 |
138   |       |
      | -3.60 |
275   |       |
      |-13.64 |
272   |       |
      |-11.45 |
97    |       |
      | -6.39 |
339   |       |
      |  0.00 |
171   |       |
      | -4.76 |
389   |       |
      | -1.59 |
383   |       |
```

```
      | -2.17 |
340   |       |
      | -1.79 |
      |-------|
INF   | -8.80 |
       -------
BEST  -325.86

13> sequence 61 233 326 320 334 203 198 31 30 51 4 211 251 254 305 127 57 105 281 138 275 272 97 339 413 171 389 383 340
sequence #5= 61 233 326 320 334 203 198 31 30 51 4 211 251 254 305 127 57 105
281 138 275 272 97 339 413 171 389 383 340

14> map
================================================================================
Map:
  Markers          Distance
   61  E32M47_8      9.1 cM
  233  E37M47_1      2.2 cM
  326  E38M48_5      5.2 cM
  320  E32M61_6      7.5 cM
  334  E33M60_5      4.4 cM
  203  E32M49_2      3.8 cM
  198  E35M48_1      2.2 cM
   31  E38M50_1      0.7 cM
   30  E38M50_1      1.4 cM
   51  E41M59_5      1.3 cM
    4  E32M49_7      2.0 cM
  211  E33M62_8      3.4 cM
  251  E36M59_1      5.5 cM
  254  E41M62_2     11.9 cM
  305  E40M62_2      5.9 cM
  127  E40M62_2      8.2 cM
   57  E37M49_7      5.6 cM
  105  E37M61_1    160.2 cM
  281  E40M47_2     55.3 cM
  138  E40M50_1     24.9 cM
  275  E40M47_1      9.1 cM
  272  E40M47_7     12.5 cM
   97  E40M47_1     20.0 cM
  339  E38M60_4     23.0 cM
  413  P10           8.5 cM
  171  E38M60_6      1.0 cM
  389  SSL29         4.2 cM
  383  SSL13         5.0 cM
  340  E38M60_5   ----------
                  403.8 cM   29 markers   log-likelihood= -336.75
================================================================================

15> sequence 323 197 29 316 359 220 77 163 123 248 76 258 346 365 324 100 358 387 345 367 33 107 78 53 377
sequence #9= 323 197 29 316 359 220 77 163 123 248 76 258 346 365 324 100 358 387 345 367 33 107 78 53 377

16> try 415

       415
       -------
      | -5.84 |
323   |       |
      | -6.43 |
197   |       |
      | -9.30 |
29    |       |
      |-12.21 |
316   |       |
      | -6.24 |
359   |       |
      | -7.54 |
220   |       |
      | -8.52 |
77    |       |
      | -5.13 |
163   |       |
      | -2.10 |
123   |       |
      | -8.23 |
248   |       |
      | -9.29 |
76    |       |
      | -8.76 |
258   |       |
      | -6.08 |
346   |       |
      | -5.94 |
365   |       |
      |  0.00 |
```

```
324   |       |
      | -1.12 |
100   |       |
      | -6.74 |
358   |       |
      | -7.06 |
387   |       |
      | -5.52 |
345   |       |
      | -5.63 |
367   |       |
      | -7.92 |
33    |       |
      |-21.06 |
107   |       |
      |-17.90 |
78    |       |
      | -9.89 |
53    |       |
      |-15.23 |
377   |       |
      | -6.66 |
      |-------|
INF   | -6.80 |
       -------
BEST  -373.20

17> sequence 323 197 29 316 359 220 77 163 123 248 76 258 346 365 415 324 100 358 387 345 367 22 107 78 53 377
sequence #6= 323 197 29 316 359 220 77 163 123 248 76 258 346 365 415 324 100
358 387 345 367 22 107 78 53 377

18> map
================================================================================
Map:
  Markers          Distance
  323  E32M61_1     25.1 cM
  197  E35M48_1      8.9 cM
   29  E38M50_7      5.4 cM
  316  E41M50_1     12.2 cM
  359  E35M62_8      5.7 cM
  220  E38M50_5      3.1 cM
   77  E41M62_2      5.6 cM
  163  E41M48_3     13.2 cM
  123  E40M62_1      5.8 cM
  248  E36M59_1      4.6 cM
   76  E41M62_1      7.4 cM
  258  E41M62_1     11.3 cM
  346  E33M50_4      7.2 cM
  365  E35M48_7     15.6 cM
  415  P25          17.1 cM
  324  E38M48_1     29.2 cM
  100  E40M47_2     17.2 cM
  358  E35M62_5     11.2 cM
  387  SSL22_2      15.9 cM
  345  E33M50_3     17.6 cM
  367  E35M48_9     60.2 cM
   22  E40M59_8     91.6 cM
  107  E33M48_3      4.8 cM
   78  E41M62_2     20.8 cM
   53  E41M59_7      8.4 cM
  377  ORS31_3     ----------
                   425.2 cM   26 markers   log-likelihood= -390.29
================================================================================

19> sequence 60 117 102 103 360 390 108 329 328 327 293 69 43 165 167 330 166 110 37
sequence #12= 60 117 102 103 360 390 108 329 328 327 293 69 43 165 167 330 166 110 37

20> try 417

       417
       -------
      | -4.06 |
60    |       |
      | -6.88 |
117   |       |
      |-12.66 |
102   |       |
      |-12.81 |
103   |       |
      |-13.27 |
360   |       |
      |-17.27 |
390   |       |
      |-19.31 |
```

```
108   |       |
      |-19.52 |
329   |       |
      |-18.41 |
328   |       |
      |-11.08 |
327   |       |
      |-12.55 |
293   |       |
      |-14.39 |
69    |       |
      |-14.53 |
43    |       |
      |-15.52 |
165   |       |
      | -9.26 |
167   |       |
      | -5.76 |
330   |       |
      | -1.80 |
166   |       |
      | -7.50 |
110   |       |
      | -3.40 |
37    |       |
      |  0.00 |
      |-------|
INF   | -4.85 |
       -------
BEST  -278.12

21> sequence 60 117 102 103 360 390 108 329 328 327 293 69 43 165 167 330 166 110 37 417
sequence #7= 60 117 102 103 360 390 108 329 328 327 293 69 43 165 167 330 166
110 37 417

22> map
================================================================================
Map:
  Markers          Distance
   60  E32M47_2     27.0 cM
  117  E40M62_7     11.2 cM
  102  E37M61_2     10.0 cM
  103  E37M61_7      8.2 cM
  360  E35M62_1      4.4 cM
  390  SSL30         3.1 cM
  108  E33M48_5      2.7 cM
  329  E38M48_1      3.7 cM
  328  E38M48_9     13.7 cM
  327  E38M48_8     10.1 cM
  293  E33M48_1      7.6 cM
   69  E36M59_1      7.0 cM
   43  E37M47_1      3.9 cM
  165  E41M48_7     12.9 cM
  167  E41M48_9     18.2 cM
  330  E38M48_1     27.9 cM
  166  E41M48_8      6.5 cM
  110  E33M48_1     12.7 cM
   37  E37M47_3     18.0 cM
  417  P45         ----------
                    209.0 cM   20 markers   log-likelihood= -278.12
================================================================================
```

Annexe J

Comparatif : Gènes Similaires d'*Arabidopsis* et Résultats Expérimentaux

Le tableau ci-dessous présente les résultats d'hybridation obtenus à partir des sondes Overgo sur la banque de clones BAC ainsi que le nombre de gènes similaires chez *Arabidopsis* aux gènes qui sont similaires aux séquences EST qui ont servi à définir les sondes Overgo.

n° sondes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
nb clone Han	0	1	5	9	0	4	3	2	6	2	4	0	1	2	3	1	5	3	0
nb gène Ath	3	6	5	10	4	2	10	1	2	2	3	1	1	9	2	3	1	2	3
n° sondes	20	21	22	23	24	25	26	27	28	30	32	33	34	35	36	37	38	39	40
nb clone Han	1	3	1	1	1	2	5	1	0	6	5	1	1	2	3	13	1	19	0
nb gène Ath	4	2	1	2	3	1	7	3	2	2	10	2	3	3	2	6	2	6	2
n° sondes	41	42	43	44	45	46	97	47	98	48	49	99	100	137	138	50	51	52	53
nb clone Han	2	2	3	2	2	0	16	1	0	0	0	0	3	0	0	2	3	2	0
nb gène Ath	2	1	1	1	1	9	4	1	2	1	2	1	3	1	2	1	1	2	2
n° sondes	54	101	55	139	56	57	102	58	59	103	104	140	105	60	61	108	62	63	64
nb clone Han	1	4	0	2	4	3	2	1	0	1	1	1	0	0	0	1	0	0	2
nb gène Ath	3	1	1	1	3	1	1	1	3	1	1	1	3	2	3	1	6	2	1
n° sondes	109	65	66	110	67	68	111	69	112	113	141	114	71	115	72	116	142	143	
nb clone Han	0	0	6	1	2	15	0	7	0	0	1	4	1	2	3	1	6	1	
nb gène Ath	2	4	3	1	4	4	1	3	1	3	6	2	2	2	3	1	1	4	

Le tableau ci-dessous présente le nombre de bandes visibles sur agarose après amplification avec les amorces sur l'ADN génomique ainsi que le nombre de gènes similaires chez *Arabidopsis* aux gènes qui sont similaires aux séquences EST qui ont servi à définir ces amorces.

n° amorces	3	5	6	7	8	9	10	11	12	14
nb bandes	1	2	1	2	5	1	1	1	1	1
nb gène Ath	1	1	1	3	6	4	1	1	2	1

Annexe K

Extraction d'ADN Génomique

Méthode modifiée d'après Fulton *et al.*, 1995.

- Prélever 70 à 80 mg de matériel végétal congelé à -20°C dans un Eppendorf de 2 mL contenant 3 à 4 billes de verre,
- Congeler dans l'azote liquide,
- Broyer au microbroyeur Restch à 5 000 cpm,
- Ajouter 750 μL de tampon d'extraction* additionnés de 2,5 μL de RNAse A (10 mg/mL, DNAse free),
- Vortexer et incuber au bain-marie à 65°C pendant 90 min minimum,
- Centrifuger 10 min à 15 000 g à 4°C,
- Prélever le surnageant et ajouter 750 μL de mélange de chloroforme : isoamylalcool (24 :1) à 4°C,
- Centrifuger 10 min à 15 000 g à 4°C,
- Prélever la phase aqueuse,
- Ajouter 1 volume d'isopropanol froid,
- Centrifuger 5 min à 15 000 g à 4°C,
- Laver à l'éthanol 70%,
- Sécher entre 10 et 15 min au speed vac,
- Reprendre dans 100 μL de TE chauffé à 65°C.

*Tampon d'extraction :

- 2,5 volumes de tampon d'extraction d'ADN [0,35 M sorbitol, 0,1 M Tris-base, 5 mM EDTA pH=7,5],
- 2,5 volumes de tampon de lyse des noyaux (0,2 M Tris, 0,05 M EDTA, 2 M $NaCl$, 2% CTAB, 1 volume de sarkosyl (N-Laurylsarcosine) à 5%(p/v),
- 5 g.L^{-1} de bisulfite de sodium.

Annexe L

La Technique de SSCP

La technique SSCP est basée sur le comportement électrophorétique d'un fragment d'ADN simple brin dans un gel d'acrylamide non dénaturant. En effet, une molécule d'acide nucléique monocaténaire peut former des structures secondaires dues à des appariements de bases au sein de la molécule. Ces structures secondaires dépendent de la séquence propre au brin d'ADN et donnent une conformation particulière à chaque type de molécule monocaténaire, (Orita *et al.*, 1989, *Proc Natl Acad Sci* USA 86 :2766 ; Orita *et al.*, 1989, *Genomics* 5 : 874). Ainsi, 2 séquences d'ADN très proches peuvent se différencier sur la base de la conformation de leur forme monocaténaire : 2 allèles d'un même gène seront distingués. Cela permet par exemple de détecter un allèle mutant (éventuellement responsable d'une maladie génétique) chez un individu. Cette technique est de réalisation simple mais comporte deux inconvénients majeurs : le comportement électrophorétique des fragments simple-brin est imprévisible (très dépendant de la température et des conditions d'électrophorèse), et pour les fragments assez grands ($>$ 200 pb) toutes les mutations ne semblent pas détectables (la méthode permet de détecter essentiellement les variations de séquence de type microinsertion-délétion).

Source : "Principes des techniques de biologie moléculaire", Denis TAGU au édition INRA (Institut National de la Recherche Agronomique, 147, rue de l'Université – 75338 Paris Cedex 07). Le gel utilisé en SSCP est un gel de polyacrylamide natif (sans dénaturant). Le haut du gel appelé "Stacking gel" sert à concentrer les échantillons avant de les séparer dans le bas du gel appelé "Separating gel".

- Monter les plaques et marquer 1 cm sous la position du peigne,
- Couler environ 1 mL d'agarose 0,3% chaud,
- Dégazer 35 mL de "separating gel[1] 9%" sous agitateur puis ajouter 17,5 μL de Temed et 87,5 μL d'ammonium persulfate 10%,
- Couler le gel avec une seringue jusqu'au trait de marquage et ajouter 1 mm d'eau,
- Laisser polymériser 30 min minimum,
- Retirer l'excès d'eau avec un papier filtre,

- Dégazer 8,75 mL de "Stacking gel[2] 6%" sous agitateur puis ajouter 7 μL de Temed et 44 μL d'ammonium persulfate 10%,
- Couler le gel à la pipette en évitant de former des bulles,
- Placer délicatement le peigne et attendre 30 min minimum,
- Enlever le peigne et secouer la plaque afin d'éliminer le liquide des puits,
- Marquer l'emplacement de chaque puits au marqueur pour faciliter le dépôt des échantillons.

- Préparer 1,5 L de tampon d'électrophorèse TBE 0.5x,
- Déposer dans un puits sur 2 quelques gouttes μL de tampon de charge dilué de moitié,
- Effectuer un pré-run de 15 min au voltage désiré.
- Reprendre les échantillons d'ADN (10 μL) dans 10 μL de tampon de charge "bleu-formamide" et les dénaturer à 94°C pendant 5 min.
- Déposer les 20 μL de chaque échantillon,
- Régler le voltage en fonction de la taille des fragments amplifiés (1 V / paire de base).
- Faire migrer toute la nuit à 4°C en chambre froide.

[1]Stacking gel 6% : 64 mL d'acryl 30%, TBE 10x 16 mL, eau 240 mL.

[2]Separating gel 9% : 96 mL d'acryl 30%, TBE 10x 16 mL, eau 208 mL.

Annexe M

Miniprep

La purification se fait sur 3 mL de culture bactérienne en milieu liquide.

- Centrifuger à 2 000 rpm pendant 10 min à 4°C,
- Eliminer le surnageant,
- Resuspendre le culot dans 300 μL de GTE[1] contenant 3 μL de RNase A (10 mg.μL^{-1}),
- Transférer dans un tube Eppendorf de 1,5 mL,
- Ajouter 300 μL de solution de lyse (0,2 M $NaOH$, 1% SDS),
- Homogénéiser en retournant plusieurs fois,
- Incuber 3 min à température ambiante,
- Ajouter 300 μL de solution de potassium 3M / acétate 5M,
- Homogénéiser en retournant plusieurs fois,
- Incuber 5 min au froid,
- Centrifuger à 12 000 rpm pendant 5 min à température ambiante,
- Transférer le surnageant dans un nouveau tube Eppendorf propre,
- Ajouter 630 μL d'isopropanol et homogénéiser,
- Centrifuger à 12 000 rpm pendant 10 min à température ambiante,
- Eliminer le surnageant et ajouter 200 μL d'éthanol,
- Centrifuger à 12 000 rpm pendant 5 min à température ambiante,
- Eliminer l'éthanol et sécher le culot au speed vac pendant 10 min,
- Reprendre le culot dans 40 μL de TE. Stocker à 4°C.

[1]GTE, Glucose Tris-HCl EDTA : voire page 198.

Annexe N

Séquences des Sondes Overgo

Les tableaux ci-après contiennent les informations relatives aux sondes Overgo. La première colonne correspond à l'identifiant de la sonde Overgo, les deuxième et troisième colonnes contiennent respectivement les séquences de l'Overgo a (OVa) et de l'Overgo b (OVb). La partie de séquences en rouge correspond à la zone de recouvrement entre l'OVa et l'OVb. La quatrième colonne contient l'identifiant du gène d'*Arabidopsis* qui présente des similitudes avec la séquence de tournesol qui a servi à définir la sonde Overgo. La cinquième colonne correspond au type de sonde Overgo définie. Les sondes Overgo e^{-20} sont définies à partir de séquences de tournesol de *fortes similitudes*, les sondes Overgo e^{-5} sont définies à partir de séquences de tournesol de *faibles similitudes* et les sondes Overgo *multi* sont définies à partir de séquences autres que tournesol (souvent *Medicago truncatula*) qui présentent des similitudes avec au moins 4 organismes différents dont *Arabidopsis* (pour plus de détails sur ces sondes, voir l'annexe *O* page 194). La dernière colonne correspond au nom du cultivar de tournesol.

n°	Sonde OV	Séquence OVa	Séquence OVb	Gène *Ath*	type	Cultivar
r1	BQ970990	GCCGTATGTGTCTTGTTGAGGTTG	GGTTTCGGAGACTTCTCAACCTCA	At5g37510	e^{-20}	RHA801
r2	BQ912790	GATCTCGTGCAAGAATTCGACAAA	GATGACCGAATCAATGTTTGTCGA	At5g37680	e^{-5}	RHA801
r3	HBK307	AAAGGTCTTCAACCGCTTCGACAC	CTTACCGTCGCCGTTGGTGTCGAA	At5g37770	e^{-5}	?
r4	BQ972149	ATGGGACCGAAAGATGTGTGGTTC	AGGCTTATATCTCTATGAACCACA	At5g37790	e^{-20}	RHA801
r5	AJ412313	GACAGGAGACTTGATGACTGCATT	GCTCCACCCAAGTAATAATGCAGT	At5g37850	e^{-5}	Emil
r6	BU022969	TGAAATGGGCACCGCAGAAGGAAG	GCAGAATGTGCCAAGACTTCCTTC	At5g38010	e^{-5}	RHA280
r7	AJ412423	GCTTCTGCTACACTGTTGTATTCC	TCAGCCGCATTTCCATGGAATACA	At5g38220	e^{-5}	Emil
r8	X05079	TCAGCCCAAGTGATGAAGGAGCTT	CCTTCTTGCATTCAGCAAGCTCCT	At5g38430	e^{-20}	undef
r9	BQ967298	GGGATAGGGATACTGTTGTTCGTG	AAAGCAGCACGAAGAGCACGAACA	At5g38470	e^{-20}	RHA801
r10	BQ916582	TAAATGGCACCCGTTTTTACTTGC	CTTTAGGTGCAGAAGGGCAAGTAA	At5g38900	e^{-5}	RHA801
r11	BU033081	CGTTATCTGGAGGGTATGAAGATG	CTCTCCATGTCCTCATCATCTTCA	At5g39550	e^{-20}	RHA280
r12	BQ976715	GATGTGCCTGAATTGCTTGATAAT	CACCCTTTACAATGTGATTATCAA	At5g40510	e^{-5}	RHA801

TAB. N.1 – Séquences des sondes Overgo définies à partir de séquences de tournesol similaires à des gènes d'*Arabidopsis* dans la région de la rubisco du chromosome 5.

ATTENTION, les séquences dont l'identifiant commence par "TC" sont des séquences dite "Tentatives Consensus". Ces séquences sont la synthèse du regroupement de plusieurs séquences EST très fortement similaires. Cependant, dans ces Tentatives Consensus peuvent se retrouver plusieurs séquences EST de cultivars différents. Il est donc très difficile de savoir à quel cultivar la séquence TC correspond.

n°	Sonde OV	Séquence OVa	Séquence OVb	Gène *Ath*	type	Cultivar
1	BQ971681	AGGCTGTAAGGCATGTTAGATCCG	CTGATGTCACCCATCACGGATCTA	At5g01410	e^{-20}	RHA801
2	BQ911224	CAAGTCTGAGGATGAGGAGCACAA	GGCCTCAACCTTCTTTTTGTGCTC	At5g02490	e^{-20}	RHA801
3	HAK346	TCGGTACCACATACTCATGTGTCG	TCGTGTTGCCATACACCGACACAT	At5g02500	e^{-20}	?
4	BQ976769	CTACAACAAGAAGCCCACCATCAC	CTGGATCTCCCGAGATGTGATGGT	At5g02570	e^{-20}	RHA801
5	AJ437714	AAGTGGTGAGGACATCAATTGCGC	ATCACAGTCAGCACCTGCGCAATT	At5g02610	e^{-20}	Emil
6	HaCot004_H01	CTGGATTTGGGCGAAAGGGTCATG	GGAATATCACCAACGGCATGACCC	At5g02960	e^{-20}	?
7	BQ975716	ACCAACAGAGGCTCATCTTTGCTG	TCCTCAAGTTGCTTGCCAGCAAAG	At5g03240	e^{-20}	RHA801
8	BQ972167	ATCATTACCCGACAGGCGAGTTTA	CATATTCAGCCACCCTTAAACTCG	At5g03290	e^{-20}	RHA801
9	BU026603	CATGTGCTGTCTGTGTTGTTGGTG	ATAAGAGACCTCTCGCCACCAACA	At5g03300	e^{-20}	RHA280
10	BQ970776	CGGATTTTGGAATGGCTCGGTTAG	AGTCCTTCTTGAGGACCTAACCGA	At5g03320	e^{-20}	RHA801
11	BU022132	ACAATGTGGTTCGGTGAGAGTGAG	TTTCACGAACGTTGGCCTCACTCT	At5g03340	e^{-20}	RHA280
12	BQ914752	AAAGGAAATACAGTGGAGAGGTCT	CAAACTAGTTCGAGGCAGACCTCT	At5g03415	e^{-20}	RHA801
13	TC1765	GAAGCGAGGATATGGAACAACGTG	TCTCAGCTCGTTCGAACACGTTGT	At5g03860	e^{-20}	*
14	TC995	GTGGTTGCAGGTATCTCGTATGCT	GGTATCCGCTGTTGATAGCATACG	At5g05170	e^{-20}	*
15	BU024951	AGCTGGCAGCATTGAAGGGTTTAG	TCCTTCAACCTTGCATCTAAACCC	At5g05780	e^{-20}	RHA280
16	TC1026	GCAATGAACTATCGGGTCGGGTTT	CGGTAGTCGCATGAAAAAACCCGA	At5g06090	e^{-20}	*
17	BU031163	GCACTTTCAATCTCGAGATAATCG	GGGTTTTTCTTCACGCCGATTATC	At5g06360	e^{-20}	RHA280
18	BU029117	ATGGTCATCCCTCATTTGACTGAG	GGGAAGCACCATAGTTCTCAGTCA	At5g06460	e^{-20}	RHA280
19	HBK181	CCACACAAGTCTAGGGAATGCCTG	GAATGAGGATCAAGGGCAGGCATT	At5g07090	e^{-20}	?
20	BU024449	CTAAGATGCTGCTTGCTGCTATTG	CCACTATGATTTTCATCAATAGCA	At5g07290	e^{-20}	RHA280
21	TC2296	ACTTTGAGCCGATTGCTGCGATTG	GTCCAGATCTCGATACCAATCGCA	At5g07340	e^{-20}	*
22	BQ915800	GGCAAGTTAGAGTGGAGGTAGATG	ACCTCAACATCAACGCCATCTACC	At5g07920	e^{-20}	RHA801
23	BQ968584	GGGCTACCTGTTTTCAACTCGGTT	CAGCTTTTGCATCTGCAACCGAGT	At5g08300	e^{-20}	RHA801
24	BU017162	GACAAGAACGCCAACTCAAGTTGC	CTTCTGAGCGTGACTTGCAACTTG	At5g08520	e^{-20}	RHA280
25	TC1934	CCTCCTCCTGAGAAAACTCACTTC	CGTCTTTCAAACCACCGAAGTGAG	At5g08530	e^{-20}	*
26	BQ971595	GATGTTTGGTCTTGGGGAGTGACA	CCAGCATCACATAAAGTGTCACTC	At5g08590	e^{-20}	RHA801
27	TC3374	ACTTGATTCCACGTCGCGTATGCT	CAAAATGTGCGGGGAAAGCATACG	At5g08670	e^{-20}	*
28	BU026456	GTAAAGCTCATGGTGGCGGAAAAC	CCCGCAAAAACACAACGTTTTCCG	At5g09670	e^{-20}	RHA280
30	BU029407	GCTTCGAACTCGAATAAGAGTTGG	CATTGGTGCAGCATCACCAACTCT	At5g09770	e^{-20}	RHA280
32	BU016153	CCTCCAGAGAGAAAGTACAGTGTC	TAGATCCTCCGATCCAGACACTGT	At5g09810	e^{-20}	RHA280
33	BQ915111	TTTAGACCGCCAAGATTATGTGCG	AGAAAGGATCTGCGCACGCACATA	At5g09900	e^{-20}	RHA801
34	TC2244	GGTAGACAAAGTGGTGGTACTCTG	TTCGGTGTTTGCAGTCCAGAGTAC	At5g10170	e^{-20}	*
35	NP524612	CCACCATTAGAGCTAGCACACCAA	CGCGACATGAGAAACATTGGTGTG	At5g10240	e^{-20}	?
36	BQ966178	ACAAGGATTCCCGATGAAACAGGG	CCCTGGTGTTAAGACTCCCTGTTT	At5g10360	e^{-20}	RHA801
37	AJ539821	GCATTGAGGGAGATCAGGAAGTAC	GTTCAGTGCTCTTCTGGTACTTCC	At5g10400	e^{-20}	Emil
38	TC3505	AGATGGGAATTTACCCACTTGTGG	GCAAGGCTTATCAGTCCCACAAGT	At5g10960	e^{-20}	*
39	TC4266	GAAGAAGCCTCATCGTTATCGTCC	GAGTGCGACAGTTCCAGGACGATA	At5g10980	e^{-20}	*
40	BU017154	TGGAACTAAAGGTCTTGCAATCAC	AGCCGAAGCCACAAATGTGATTGC	At5g11170	e^{-20}	RHA280
41	BQ910003	GAAGGGCACAAGAGAATCTTGGTG	CGACTGAATCTGTTGCCACCAAGA	At5g11200	e^{-20}	RHA801
42	BU018273	CATGCGAAGTTGCAGGCGCTTAAG	TCTCCGCTTCTTTAGCCTTAAGCG	At5g11230	e^{-20}	RHA280
43	BU024722	TCTACTTCAAAGCCCGACCTGAAG	ATGGTGTAGTCTCCAGCTTCAGGT	At5g11500	e^{-20}	RHA280
44	BU024438	CTTTCACAGGGCTTAATACTAAGC	GTCATGAAGGTCACTTGCTTAGTA	At5g11520	e^{-20}	RHA280
45	BU032937	GTTGCTGGTACTCTTACCAACAAG	TCAAAGCAGGAGCCATCTTGTTGG	At5g11770	e^{-20}	RHA280
46	HaCot002_F10	CATTGGTACACAGGTGAGGGTATG	TGAACTCCATCTCGTCCATACCCT	At5g12250	e^{-20}	?
47	BQ968778	CTCACATTGGTGCCATGTTCACTG	GCAACCGACAAAAAAGGCAGTGAAC	At5g12860	e^{-20}	RHA801
48	BU036267	GTGAATTGTAGGAACGGAATGCCA	TTGGGTGCAAATGGCATGGCATTC	At5g13010	e^{-20}	RHA280
49	BQ968448	GTGGTGACAACATACGAAGGAGTG	CGATTGGATGTGAGTGCACTCCTT	At5g13080	e^{-20}	RHA801
50	TC1101	AGATGAAATATCCGGATTATGCTG	AGTCCTGAGCTTGTCTCAGCATAA	At5g13410	e^{-20}	*
51	BU025805	CTGTTGAGGTTTCACCTCGGCTTG	CCTTGGGTATCATCAGCAAGCCGA	At5g13420	e^{-20}	RHA280
52	BQ916761	AAAGACGTTCTTGCCCTTGCTTCC	AGAGGTCAACTTCGAGGGAAGCAA	At5g13430	e^{-20}	RHA801
53	BU020314	TTGCATCCCGTTACCAAATGCTGG	CCACCCACCAAAATCACCAGCATT	At5g13440	e^{-20}	RHA280
54	BQ911963	GAGGCTGTGAAGTACAAGAACACG	GACTGAAAGCATCAAACGTGTTCT	At5g13490	e^{-20}	RHA801
55	HaCot001F_F07	ATGAACACCAACCCGAACTCCTTC	TCTGCAGCAACTTCCTGAAGGAGT	At5g13630	e^{-20}	?
56	AJ412627	CTGTTCGTGATTTCAAAGCCTGAT	CGGGACTTTTAAACACATCAGGCT	At5g13850	e^{-20}	Emil
57	AJ412307	AGCAAAGATTTGGGAGTGAGATTC	TCGGTTGGTTGAATGGGAATCTCA	At5g13870	e^{-20}	Emil
58	BU026384	CATCTGGATTTGGGGTATTGCTCG	TTGAGTCCTTGCTCCTCGAGCAAT	At5g14040	e^{-20}	RHA280
59	BQ969663	CTGGATTAAGGATCCTGCAATTCC	GAAGCCAGTGACAATAGGAATTGC	At5g14060	e^{-20}	RHA801
60	BQ917029	AATATTCGGGGCTACCGAGTACAC	GACATCGATTGCTGTGGTGTACTC	At5g14640	e^{-20}	RHA801
61	BU015936	CATGCATGTTTGATGCAGGTGATC	TTTCTCGAACGTTGCCGATCACCT	At5g14720	e^{-20}	RHA280
62	AJ541540	AGTCTTCTCGGTGAAGGGGGTTTT	CTTTAAACACACACCCAAAACCCC	At5g15080	e^{-20}	Emil
63	BQ912180	TTCAAGAAGCCACGACGTCCTTAC	CCAAACGTTCCTTCTCGTAAGGAC	At5g15200	e^{-20}	RHA801
64	BU027548	CTTTGGCACCATTTGGTGGGGTTT	GATGAGATTAAGGGCGAAACCCCA	At5g15410	e^{-20}	RHA280
65	BU025750	GCCTTCTTGGCACAAAGAATCTCG	ACATGGCGTTCACAGACGAGATTC	At5g15490	e^{-20}	RHA280
66	BU018657	TGGAGCTTCCTGAATGGACTGATA	GTAGCAGTCTTCACAATATCAGTC	At5g15520	e^{-20}	RHA280
67	BQ967241	CCGAGAAAGGTCTACTTCAACGGT	GCAGCATGCATTCTTCACCGTTGA	At5g15630	e^{-20}	RHA801
68	AJ412408	ATTTCGACCGTGAGCTAATCGGAC	CCGAAGTACATTGCTGGTCCGATT	At5g15650	e^{-20}	Emil
69	BU024718	GGCTGCATATACGAGTGGATGGAA	GGCAGTTTTCACTTCTTTCCATCC	At5g15790	e^{-20}	RHA280
70	BQ965540	GCGCTTTAGAGTTTGGAAAAGCTG	GCTCCTTTTCTGGTTAGCAGCTTTT	At5g16000	e^{-20}	RHA801
71	TC2775	CACTTGATGGAGAAGATTGAATCC	GGTGCAACAAGTTCTCGGATTCAA	At5g16440	e^{-20}	*
72	BQ911917	AGAGTTGGGTTTGAGGCACTCTGA	GTACGCAGCAATGTGTTCAGAGTG	At5g16570	e^{-20}	RHA801
73	HaCot009_D10	TAGCAGCAGTGCTTCAAGGACTGT	CCCGCAACCACTTTTAACAGTCCT	At5g16910	e^{-20}	?
74	TC3123	GTTCTTCCTGGGAATCTTGGCTTG	CATAAGCTTCATTCAACAAGCCAA	At5g17230	e^{-20}	*
75	HaCot007_G05	ACCACTTGGTCCAGAACAAGAATG	ACCTCCATGCAGTACTCATTCTTG	At5g17310	e^{-20}	?

Suite du Tableau page suivante...

n°	Sonde OV	Séquence OVa	Séquence OVb	Gène *Ath*	type	Cultivar
76	BU027110	CTATTCATGAGGACGCTGCAAGTG	GGTGCAATAAAACCTCCACTTGCA	At5g17330	e^{-20}	RHA280
77	BU016722	AGTGGAGTCAGCATAGAAGAATGG	ACTGCTCGTTTCTCCACCATTCTT	At5g17420	e^{-20}	RHA280
78	BU025411	CTAGAAAACCCAAAAGACATGACC	AGATGAGATGCACCTGGGTCATGT	At5g17770	e^{-20}	RHA280
79	TC541	AGAAGGATTGCCCCTTAGGAAAGC	GTAGAAAGCGTGTTCGGCTTTCCT	At5g17920	e^{-20}	*
80	BU027560	CCCTGACATATTTGCAAATCCAGG	ACTGACTGTTACCCCTCCTGGATT	At5g18170	e^{-20}	RHA280
81	BQ917146	CATTGCCGGTGCTAATAACGGAAG	CATCACCACAAGAGAGCTTCCGTT	At5g18190	e^{-20}	RHA801
82	BU672036	GGTAAGGTATGACAGGACTCTGCT	TCTCGGATCAGCAACAAGCAGAGT	At5g18380	e^{-20}	RHA280
83	HAK142	GCCAAAGTCTATCGGTTTGCTTCC	TACACCGCCATCAAACGGAAGCAA	At5g18410	e^{-20}	?
84	TC2504	GGATGGGTTTTGGTGATAATACTC	GGCCAGTAACCAACTAGAGTATTA	At5g18460	e^{-20}	*
85	BU026743	TGATGCTGATCTGCTCAAAGGTGC	GAGTGACTTGCCACATGCACCTTT	At5g19140	e^{-20}	RHA280
86	BQ967098	TCACGCACGCGTTATTTTACCTGG	AAAGCATGGCGTCATCCCAGGTAA	At5g19180	e^{-20}	RHA801
87	HO0008G07FT	ACCGACATGAAGAAGCTAGAGGAG	CAACACTACGAACAGCCTCCTCTA	At5g19510	e^{-20}	?
88	HaCot004_B01	ATGAGTGGACGGTTGAGCTTAAAG	ATCCGATCAGCCATTGCTTTAAGC	At5g19550	e^{-20}	?
89	TC1576	TATGCCTCTGAGGGATGGTTCATG	TACCCTCCATAAGCTTCATGAACC	At5g19670	e^{-20}	*
90	BQ965817	CCTGTCTGTGGAAACTGTGTGTGT	TACGGGGGCAGTAAATACACACAC	At5g19690	e^{-20}	RHA801
91	BU025035	TTCAGCGAAACTGGGTCAGGAAAG	TCGCTCTTGGAACATGCTTTCCTG	At5g19770	e^{-20}	RHA280
92	HaCot005_F08	CTTCTTGAAGCGTGCAATGACGAG	GTCGAACATCTGGACTCTCGTCAT	At5g19820	e^{-20}	?
93	BU034616	GAACCGAGTGGATGATCAGGATAC	AAGTTGCGCGATATGCGTATCCTG	At5g19900	e^{-20}	RHA280
94	BU025959	AACCATTGGGGTTGAGGTTCATCC	GGTGAAGAAATCCAAGGGATGAAC	At5g20010	e^{-20}	RHA280
95	BQ965467	ACCATTGGTGTTGAGGATCATCCC	GGGTGAAGAAATCCAAGGGATGAT	At5g20020	e^{-20}	RHA801
96	BQ967471	AAAGCTTGGAAGAGGCTGTGAGGA	TCCTCCCATGTTTCTCTCCTCACA	At5g20070	e^{-20}	RHA801
97	BU027121	GGTGAAACCGTTAAAGCAAGACTC	TCACTGAGAATTGCTTGAGTCTTG	At5g12480	e^{-20}	RHA280
98	CD855185	TCAAGGCCATTTGAGTATCTGAGG	CACCTAAGCTAGTGAGCCTCAGAT	At5g12980	e^{-5}	psc8
99	CD849118	CGATGAAACGCAGAACCCTTGTAG	GAGAGAAAGGTTTTGGCTACAAGG	At5g13120	e^{-20}	psc8
100	CD852272	GATTGCCTCCCGGTTTCAGGTTTC	TCTTCATCTGTTGGGTGAAACCTG	At5g13180	e^{-5}	psc8
101	BU034867	CAAGAACGATTCATTTTCATGAGA	CGTTGTAAGCGGTTTCTCTCATGA	At5g13580	e^{-5}	RHA280
102	BQ968074	GTACTCGTTGTCTGCTCCGAGATC	GGAACGTTACCGCAGTGATCTCGG	At5g13930	e^{-5}	RHA801
103	CD854117	AATGTCACTGTCTTTGAAGCTGAT	TCCCGCCTACTCTCCCATCAGCTT	At5g14220	e^{-5}	psc8
104	BU024965	CTCCGTGATGAAGTCTCCAAGTAC	GGTCTCCTTCAATCATGTACTTGG	At5g14320	e^{-5}	RHA280
105	BQ968141	ATTACGACCTGGGTCTTCCTAACC	TCGTCAGTAGCATCACGGTTAGGA	At5g14590	e^{-5}	RHA801
106	CD855634	ATCGAAGCTTTGGCTGATGGAGCT	GTAAACCTGCAGCTACAGCTCCAT	At5g14800	e^{-5}	psc8
107	BQ974976	CAGATATCCGTGGTGAATGGATTG	TGTGGTGCGATATCAACAATCCAT	At5g14900	e^{-5}	RHA801
108	BQ976133	CCATCAACAACAGGAGACTTCCTG	ATGCACATGCTGAAGCCAGGAAGT	At5g15070	e^{-20}	RHA801
109	CD847842	GGAGTCCCAAGTACATGTCATTGC	ATGCGAAGATGATTGAGCAATGAC	At5g15470	e^{-5}	psc8
110	CD854754	GAATACATAAGAGCTGTTGCTCCA	GATCTTCTTCCTTGCGTGGAGCAA	At5g15550	e^{-5}	psc8
111	CD851096	AGCTGTGACATATATAGAACAAGG	TCCCACACGGACGTGTCCTTGTTC	At5g15750	e^{-20}	psc8
112	CD852629	GATCTGACAAATTGTCCCTTTAGG	TTATCTCGTCCACCACCCTAAAGG	At5g15920	e^{-20}	psc8
113	CD854473	GGCAGTAGTGATTCATGTTCCTTA	AGCTTTCCTCAATCTGTAAGGAAC	At5g16130	e^{-5}	psc8
114	CD845622	CGTTTTGGCCGATTGTGAATGCAG	ACAGTCTTATCACCGGCTGCATTC	At5g16400	e^{-20}	psc8
115	CD850376	GCGGTCACGCAAAGTTCGAATGTC	GTGATCTTGCAGTGAGGACATTCG	At5g16470	e^{-20}	psc8
116	BQ967059	TGGAAGCTTTGGAGAAGATGATGG	TGTACAGTCGGATCCTCCATCATC	At5g16620	e^{-5}	RHA801
117	BQ911959	TGATGAGCTTTGGCCTGCTGACAG	TTTGGTGAACCCAAGCCTGTCAGC	At5g17170	e^{-5}	RHA801
118	CD849026	CTAACCATTTCGGGTATCTTTCTC	CCCATGAAAGAAATCGGAGAAAGA	At5g17550	e^{-5}	psc8
119	BU031700	ATGTATGGGGGGATGGAGTTGTCC	TGAGCCGAAACTTCTGGGACAACT	At5g17670	e^{-5}	RHA280
120	BQ966035	AAAGAGTAGGGAAGGCGCGGAAGA	TACAACAAGTACCCTCTCTTCCGC	At5g17760	e^{-5}	RHA801
121	CD846759	GTTGTGAAGCTAAGGGTGCTATCC	GAGCATGTGGCACAAAGGATAGCA	At5g17840	e^{-5}	psc8
122	CD853282	CCGTTTCAACGAGGATATACTGAG	ATTGCGGTTCCATACGCTCAGTAT	At5g18110	e^{-5}	psc8
123	BU025316	TGGCTGCCAAAGAGTTAAAGAAGC	TGGTACCTCCATGATTGCTTCTTT	At5g18230	e^{-5}	RHA280
124	CD851336	ACTTGGGTATGGCAACTGGGACGA	AAATGCTGCCTTCAGCTCGTCCCA	At5g18620	e^{-20}	psc8
125	AJ437808	TTCAAAGGTACCCACGGCCTGAGA	AACGATTGATACTCCTTCTCAGGC	At5g19420	e^{-5}	Emil
126	BU026153	GGTGCTCGTGATGATAAGGGAAC	AGTCTCTTATCATAGGGTTCCCTT	At5g19620	e^{-5}	RHA280
127	CD854502	TCCAAGGCATCAAACACACTCAAG	TAGAGACATAGAGTTCCTTGAGTG	At5g19680	e^{-20}	psc8
128	BU027791	CTGTCGTAAGAGCGATGGCTTTGA	GCAAGCATTCCCATGTTCAAAGCC	At5g19760	e^{-20}	RHA280
129	CD845885	CTTCATAACCTTGGCGACTGCCAT	CCAGGAAGATTTTGAGATGGCAGT	At5g20000	e^{-20}	psc8
130	CD848648	ATAGTTACGAGGCTGCACCATCCA	ATTCATGAGGTTTGCGTGGATGGT	At5g20090	e^{-20}	psc8
131	HaCot010_H06	TCACTCGTGCTGTCATCGCATGTT	TCGTTACTGGTCACTGAACATGCG	At5g20160	e^{-20}	?
132	CD854509	CTCAGACACTTGTCAAGAGTGCAA	GCGTCAACCTGAACAATTGCACTC	At5g20290	e^{-20}	psc8
133	BQ914678	GAGGTAAACGATCCTGCTTTGCTA	GTGACCAATTACCTGCTAGCAAAG	At5g20350	e^{-20}	RHA801
134	U94782	GAACTTAGTCCCCATGTTTTTGCC	ATGCGACATCTGCTACGGCAAAAA	At5g20490	e^{-20}	?
135	HaCot011_G12	GTCAAGCTAATCAGGCTAGCGCTA	GTACACTCTTCGCTAGTAGCGCTA	At5g20570	e^{-20}	?
136	X14333	AGCTCTGACACCATTGACAACGTC	CTTGGATCTTGGCCTTGACGTTGT	At5g20620	e^{-20}	?
137	BF646099	GCAGTTAAAGCTCACCAGTTTTTC	CCCATGATTCCTCTGTGAAAAACT	At5g13240	*multi*	Mtr
138	BE325988	GAGAGGCATATGATGGAGACATTG	AGTGCACTTGCAAATGCAATGTCT	At5g13300	*multi*	Mtr
139	CB891199	GCGACAAGCTTTTATGAATTTGGC	GGAAAGATTCCCCCCAGCCAAATT	At5g13710	*multi*	Mtr
140	CA921615	GCTTTCTCCTCTCCTTGCTTTCTC	AGGTGAAAAGATGTATGAGAAAGC	At5g14460	*multi*	Mtr
141	U70374	TCATTTCAAGAGCGTGCCACCTTC	TGTTTCCGTGAGAGATGAAGGTGG	At5g16230	e^{-5}	?
142	CD845949	TATTAGCCCAGAAGGTAAAGTCCC	ATCAACCTTCGCCACTGGGACTTT	At5g16705	e^{-5}	psc8
143	BQ916240	GAAAGTGGATCCAGACAGATGGGA	CCCTTCGTTTGCAAATTCCCATCT	At5g16820	e^{-5}	RHA801
144	CA922382	GAATAGTTGTTTTTGCCAGCAGCAT	GTGATGCTTGGTCTGGATGCTGCT	At5g17060	*multi*	Mtr
145	AW774297	CTGCCACATTGCCTCTGTATGCAC	AGGACCTATTTGTTCTGTGCATAC	At5g18580	*multi*	Mtr
146	BG453188	TTGGAGGACCAGGGAAGGCATTGA	CCTTGTTCCAATGGAGTCAATGCC	At5g18660	*multi*	Mtr
147	AY108560	TGATTGCTGACGCTATTGACAAAG	ACACCGTCTGGTCCAACTTTGTCA	At5g18820	*multi*	Zma
148	BG585673	AGAGCAGCATCATTGTTTCCAGCT	GAGACAGACAGCTTTCAGCTGGAA	At5g19000	*multi*	Mtr
149	AW585835	CTGCGCGAGATAAAAACTTCTTAGA	CAGGATGTCTTAGAAGTCTAAGAA	At5g19010	*multi*	Mtr
150	BU019525	GTGTTCGTGTATATTGGAACAACT	CAGAAAGGTCCTTGAGAGTTGTTC	At5g19070	e^{-5}	RHA280
151	CA923153	CTTCCCATCTAATATTTGGGATCT	GATGCAAGGGACATAGAGATCCCA	At5g19330	*multi*	Mtr

Fin du Tableau

TAB. N.2 – Séquences des sondes Overgo définies à partir de séquences EST similaires à des gènes d'*Arabidopsis* localisés dans la région 0-7 Mbases du chromosome 5.

Annexe O

Définition des sondes Overgo *multi*

Les séquencs EST ont été alignées avec la séquence codante des gènes d'*Arabidopsis* afin de définir des sondes Overgo les plus conservées possibles.

```
ath_At5g13240  : GCtGTcAAAGCaCAtCAGTTTTTCtCgGAGGAaagcTGGG
bra_CA991500   : GCtGTgAAAGCaCAtCAGTTTTTCtCgGAGGAgagcTGGG
mtr_BE941418   : GCaGTtAAAGCtCAcCAGTTTTTCaCaGAGGAatcaTGGG
Sondes 137     : GCAGTTAAAGCTCACCAGTTTTTCACAGAGGAATCATGGG

ath_At5g13300  : GcGAgGCaTAcGATGGgGAcATTGCtTTcGCAAGtgCaCA
mtr_BE325988   : GaGAgGCaTAtGATGGaGAcATTGCaTTtGCAAGtgCcCA
osa_AK073117   : GgGAaGCaTAtGATGGaGAtATTGCaTTtGCAAGttCaCA
osa_AK121123   : GaGAaGCtTAtGATGGaGAtATTGCaTTtGCAAGctCaCA
Sondes 138     : GAGAGGCATATGATGGAGACATTGCATTTGCAAGTGCACA

ath_At5g13710  : GctActAGCTTtTAtGAgTacGGaTGGGGaGAaTCcTtCC
mtr_CB891199   : GcgAcaAGCTTtTAtGAaTttGGcTGGGGgGAaTCtTtCC
osa_AK072893   : GccAccAGCTTcTAtGAgTatGGtTGGGGtGAaTCcTtCC
zma_AF045570   : GccActAGCTTcTAtGAgTatGGtTGGGGtGAaTCcTtCC
ppa_BJ192576   : GttAacAGCTTcTAcGAgTatGGgTGGGGaGAgTCtTaCC
Sondes 139     : GCGACAAGCTTTTATGAATTTGGCTGGGGGGAATCTTTCC

ath_At5g14460  : aGGTGAgAAgaTGTAcGAgAAgGCAAGaAgaGGaGAAAcc
osa_AK065542   : tGGTGAaAAaaTGTAtGAcAAaGCAAGgAggGGtGAAAca
zma_AY108739   : cGGTGAaAAgaTGTAtGAcAAaGCAAGgAagGGtGAAAca
mtr_CA921615rv : aGGTGAaAAgtTGTAtGAgAAaGCAAGgAgaGGaGAAAgc
Sondes 140     : AGGTGAAAAGATGTATGAGAAAGCAAGGAGAGGAGAAAGC
```

```
ath_At5g17060  : GTtATGCTgGGgcTgGAtGCtGCtGGaAAaACaACtATtc
mtr_CA922382rv : GTgATGCTtGGtcTtGAtGCtGCtGGcAAaACtACtATtc
mtr_BF649211   : GTcATGCTtGGtcTtGAtGCtGCtGGaAAgACaACtATat
ppa_BJ197722   : GTgATGCTtGGatTgGAcGCgGCaGGcAAaACcACcATtc
Sondes 144     : GTGATGCTTGGTCTGGATGCTGCTGGCAAAACAACTATTC

ath_At5g18580  : cTGtCAcATtGCctCtGTaTGcACtGAacAaATcGGtCcc
zma_AY103889   : cTGcCAtATcGCcaCaGTcTGcACtGAgtAgATtGGtCag
mtr_AW774297   : tTGcCAcATtGCctCtGTaTGtACaGAacAaATaGGtCct
osa_AK070081   : tTGcCAtATtGCaaCaGTaTGcACaGAgcAgATcGGcCag
Sondes 145     : TTGCCATATTGcCACAGTATGCACTGAACAAATCGGTCCG

ath_At5g18660  : TTGGtGgacCaGGcAAGGCaTTaACgCCaTTaGAgCAAGG
bra_BG543436   : ----cGcgtCcGGgAAGGCcTTgACgCCtTTgGAgCAAGG
mtr_CA919243rv : TTGGaGgacCaGGgAAGGCaTTgACtCCaTTgGAaCAAGG
Sondes 146     : TTGGAGGACCAGGCAAGGCATTGACGCCATTGGAGCAAGG

ath_At5g19000  : aGAgCAGCAtCAtTGTTtgCaGCTaAAaGCaGTgTGtcTC
bra_BG543221   : gGAgCAGCAtCAtTGTTtcCaGCTaAAgGCtGTcTGtcTC
mtr_BG585673   : aGAgCAGCAcCAtTGTTtcCaGCTgAAaGCaGTcTGttTG
osa_AK073258   : tGAaCAGCAcCAcTGTTatCaGCTaAAaACtGTtTGccTG
zma_AY104286   : tGAaCAGCAcCAcTGTTatNgGCTaAAaACtGTtTGctTG
Sondes 147     : AGAGCAGCACCATTGTTTCCAGCTAAAAGCTGTCTGTCTG
```

Annexe P

Pattern de dépôt des clones BAC positifs

L'inconvénient des hybridations avec peu de sondes Overgo (moins de 5 sondes en même temps) est d'avoir un trop faible bruit de fond qui rend la distinction des champs plus difficile. Pour faciliter la lecture des membranes, celles-ci ont été faites avec un pattern spécifique contenant à la fois un clone BAC qui servira de témoin positif d'hybridation (le clone BAC $149P10$ spécifique de la rubisco) et 6 clones BAC qui serviront de témoins négatifs d'hybridation ($1A12$, $1B13$, $1M04$, $1F03$, $1H12$ et $1C03$ qui ne s'hybrident avec aucune sonde).

Ces membranes sont hybridées avec deux sondes, la sonde Overgo Rub2 (aussi notée r8, spécifique de la rubisco) et une autre sonde correspondant à l'une des sondes utilisées en pool. L'utilisation de 2 sondes diminue le bruit de fond de la membrane ce qui accroît la difficulté de lecture, mais les clones BAC spécifiques de la rubisco s'hybrident avec la sonde Rub2 et permettent ainsi de distinguer les extrémités de la membrane.

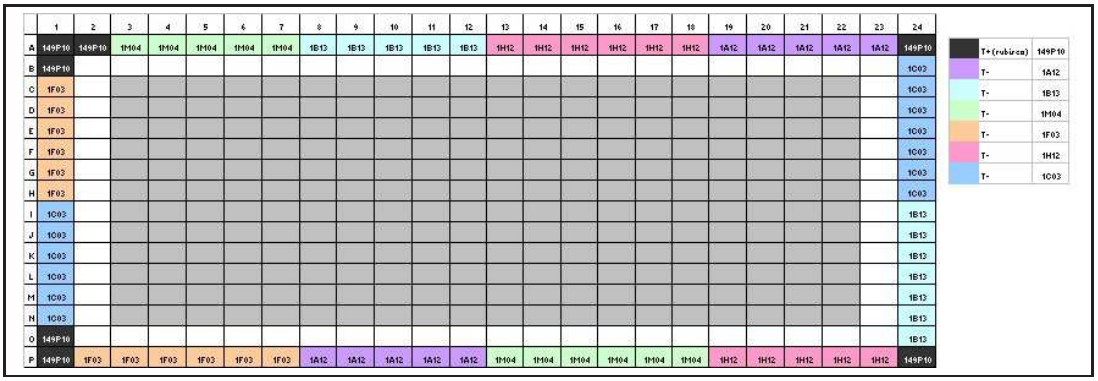


FIG. P.1 – Pattern de repiquage des plaques de clones BAC positifs qui serviront aux membranes.

Les clones BAC positifs sont repiqués dans 2 plaques 384 suivant le pattern de la Figure P.1. Le clone $149P10$ spécifique de la rubisco sert de témoin positif d'hybridation et est placé

aux extrémités de la plaque. Le coin en haut à gauche (puits : A01, B01, A02) contient 3 clones qui permettent d'orienter la membrane. Dans les colonnes 1 et 24 et les lignes A et P (c'est à dire les lignes et colonnes qui font le pourtour de la plaque) sont disposés des clones BAC qui serviront de témoins négatifs d'hybridation (1A12, 1B13, 1M04, 1F03, 1H12 et 1C03). Les lignes B et O et les colonnes 2 et 23 ne contiennent aucun clone BAC (témoin de contamination). La zone grise au centre de la plaque contient tous les clones BAC positifs. Un total de 503 clones BAC positifs a été repiqué dans 2 plaques 384 suivant ce pattern.

Ces 2 plaques 384 ont été déposées sur des membranes 11x7,5 (format plaque 384). Le motif de dépôt 1 correspond aux clones de la plaque 1 et le motif de dépôt 2 correspond aux clones de la plaque 2.

1	2
2	1

TAB. P.1 – Motif de dépôt des plaques repiqués sur les membranes 11x7,5.

Annexe Q

Extraction d'ADN de Clone BAC

Les bactéries utilisées pour la banque BAC sont des *E. coli* ayant un vecteur de type pBeloBACII résistant au chloramphenicol.

- OverNight à 37°C les clones BAC dans 5 mL de LB+chloramphenicol (12,5 mg.L^{-1}),
- Centrifuger à 2 000 rpm pendant 10 min à 4°C,
- Resuspendre le culot avec 300 μL de GTE[1] (+ 3 μL de RNAse A [10 mg.mL^{-1}]),
- Transférer dans un tube Eppendorf de 1,5 mL,
- Ajouter 300 μL de solution de lyse[2],
- Retourner doucement 5 à 6 fois et incuber 3 min à température ambiante,
- Ajouter 300 μL de 3 M acétate de potassium[3],
- Remuer doucement et incuber 5 min dans la glace,
- Centrifuger 5 min à température ambiante ou à 4°C à 12 000 rpm,
- Transférer le surnageant dans un tube propre,
- Ajouter 630 μL d'isopropanol froid,
- Centrifuger 10 min à 12 000 rpm à température ambiante,
- Laver le culot à l'éthanol 70%,
- Sécher entre 10 et 15 min au speed vac,
- Reprendre dans 50 μL de TE.

[1]Solution GTE : Glucose Tris-HCl EDTA (50 mM Glucose, 25 mM Tris-HCl pH=8.0, 10 mM EDTA pH=8.0).

[2]Solution de lyse : 0,2 M $NaOH$, 1% SDS.

[3]Solution 3 M acétate de potassium : 60 mL acétate de potassium 5 M, 11,5 mL acide acétique glacial, qsp 100 mL.

Annexe R

Séquences des Couples d'Amorces

Le tableau ci-dessous contient les informations relatives aux différents couples d'amorces définies. La première colonne correspond au numéro du couple d'amorces. La deuxième colonne contient l'identifiant du gène d'*Arabidopsis* qui présente des similitudes avec la séquence de tournesol qui a servi à définir le couple d'amorces. La troisième colonne contient l'identifiant de la séquence de tournesol qui a servi à définir le couple d'amorces. Les quatrième et cinquième colonnes contiennent respectivement les séquences de l'amorce Forward et de l'amorce Reverse. La dernière colonne correspond au nom du cultivar d'où provient la séquence qui a servi à définir le couple d'amorces.

n°	Gène *Ath*	Seq. EST	Amorces Sens (Forward)	Amorces Anti-Sens (Reverse)	Cultivar
1	At5g01160	CD852222	AGCATGCTTTTTGCCTTGAT	AGATGCCTTCCATCATTTTGA	psc8
2	At5g01530	CD847492	TTTAGGGAGTGCGAGCTCAT	ATGGAAGTGGTTGTCCAAGG	psc8
3	At5g03415	BQ914752	GGTCTTCAAAATCTCATCAAACG	TCGAAATGAACAAGCTGCAT	RHA801
4	At5g04885	BQ969550	GCAATACAAGGTGCAAAAGC	TGACGCTGTTCATGGTCATA	RHA801
5	At5g06360	CD850855	AGAAATGGATATCGCCCTGA	TTTTTCATTTGAGCTTTCTCTGC	PSC8
6	At5g07920	BQ915800	TGTGATGCAAAGGTTGCTCT	CAACGCCATCTACCTCCACT	RHA801
7	At5g08670	BU025288	AAATGAATGAACCCCCAGGT	GACGGCAGATGGAATACGAC	RHA280
8	At5g10980	CD855786	ATGGCTCGTACCAAGCAAAC	ACGAAGGGCTACGGTTCC	psc8
9	At5g12480	BU027121	AACCCGAAAACTTCTTGTTTGA	GGATCACACCAGCACTCCAT	RHA280
10	At5g14220	CD854117	TCACTGTCTTTGAAGCTGATGG	TGTTTATCTCGAAGCCCAAGA	psc8
11	At5g15070	BQ976133	TTCGGAACAGTTCTTTCATTCC	ATTATTTCATTGAAGAAGAAGATTTCG	RHA801
12	At5g16400	CD845622	GATACGTTTTGGCCGATTGT	GCAATCAAGCTTTAGAAAGACCA	psc8
13	At5g17920	BQ967599	AAAGCCGAACACGCTTTCTA	TGACATCAGCATCCCATGTT	RHA801
14	At5g19180	BQ967098	CAAGAACCGCAGCTCACTG	CCACTGCATATGTTCAGGGTTA	RHA801
15	At5g20290	CD854509	CACTGGTGGCAAGAAGAAGG	ACCGGTATCCAACCTTAAAGC	psc8
16	At5g22360	BQ914824	TTTTCAAGGGTTTTGCATCA	CTCAATCCGGTCACCTCTGT	RHA801
17	At5g23540	BQ966463	TGAAATGGTGGTTGGATGG	TGGGGGTTAATTAAACGGAAG	RHA801
18	At5g24300	BU027679	AAATTCCGTGGATGGGTTG	AACCACATGGCTCGAATCTC	RHA280
19	At5g25220	CD845764	CACATTATGTTCTGTTGCTTTGTTC	CAGACATTGTTGCCCCAGT	psc8
20	At5g26751	BQ966740	TATATGGCAGAGCGTGTCGT	TTAGGATGATCAAGAAGCCTCA	RHA801
21	At5g27850	CD855119	AGCAAACGCACCAAAAGAAC	GAAATCGGAGGCTTGTTGAC	psc8
22	At5g28740	BQ968201	TCCCCACGTGAAAGACATTT	CATAAACCCTCATGGCTCGT	RHA801
23	At5g30510	CD857284	TGGAGTCGGTTCCAGTTTCT	GGAATCAACGGTCTTCTCCA	psc8
24	At5g32451	BQ977573	CAAACGCTCCGTTTAACCA	AAAAGTGAAATCCGTCGATCA	RHA801
25	At5g35360	CD851119	TGGCTACATTGGTGTTGGTACT	CGGAGTTTCTCACCCATAGC	psc8
26	At5g36290	BQ972005	GTAGCAGCACTGTTTGTGTGC	AATAGCAGCTCTTATGGCAATG	RHA801
27	At5g37720	AJ542381	TCTCTGAGATTGGGGAGCTG	GCATCACTTCTTCTAGCAAACAA	Emil
28	At5g38430	CD846844	GTCGACTACTTGCTACGCAAAA	CATCACTTGGGCTGAATCG	psc8
29	At5g39740	CD854257	ATTGCACTGGGCTTCTTTTG	CCAAAAACACGGTTTCCAGT	psc8
30	At5g41000	BQ979503	TCCACATGAATCTCCGATACC	GCCACAGCGATGTTAATCAAT	RHA801
31	At5g42270	BU024570	AAGGCTCGAGTCCGGTTTAT	CTCGCCACACGAGAAACTT	RHA280
32	At5g42970	CD849985	ATGTTTGTGTACAGTTTGCTTGC	AGAGAGAATCTTGAGGAAACCTGA	psc8
33	At5g44030	BQ966800	AGAAAACTGAATGGGGCAAA	ACCGATTTCCATCCTCTACAAT	RHA801
34	At5g45775	CD854770	ACACTGGCTGCTTTGGTTTT	GATACCAACACGAGCCTTGC	psc8
35	At5g46630	CD849255	CTGGCCTGGAAATTTTCGTA	AAAGAGCGTTTTTGGTGCAA	psc8
36	At5g47770	AF019892	ATCGCCGGATTGTTCAGTA	TTCAACATGATTGTCCAAATCC	?
37	At5g49460	BU025367	CCACAAATCCCTGAGGATCTTA	ACCGACACCCATTCCTTTTT	RHA280
38	At5g50460	CD850857	CGTCTAGTCAAGAGATGCCACA	AAAAACCCTACAAATCCCATCA	psc8
39	At5g51400	BQ910266	AGGTGTGTTGTTCGGTTGGT	CCCTTCATCTCCCTGTGTTC	RHA801
40	At5g52920	BQ967817	GTTGCAAGAGGGGACCTTG	CACGCGTAGGTGTAGGATGA	RHA801
41	At5g53970	BQ969028	GACGACATCGATTTCTGTTTCA	TCTGCTGCAAAAGTTATCCTCA	RHA801
42	At5g55310	BU025677	ATTGGGCACTTCGAGAATCA	CAGAATCTGAAACTAGGATCCACA	RHA280
43	At5g56670	CD853019	AAGGTTCACGGATCGTTGG	CCCCCTCTTCTTACCGAAAC	psc8
44	At5g58030	AJ542335	TAGTTGTTCGAGGTCGTTGTTG	TTATTTCAGTGCCGAAAGACG	Emil
45	At5g59410	CD851706	AAGAATTCGGTGGTCATGGA	AAGTGTGCCAAGCTGTAAACAA	psc8
46	At5g60690	BU027098	GCTTGAAGTTATTAGACTCGAAGGTC	CATCTCATCAATTGGAGCAAAA	RHA280
47	At5g62020	CD846305	TTCGCTAAGGATTTGCTTCC	CCCAGCGATCAGGTACAACT	psc8
48	At5g63890	BQ917222	AATCACCAAGCAATGCAACA	ACAGACCCTGCGTTTTCAAT	RHA801
49	At5g65000	BU026233	TTCCGGTCTTGATAGCTTCTG	TCACCATCTGGAGATTTGTACG	RHA280
50	At5g66240	CD853885	CAATTAATAGCAAAAAGTGTGGGATT	GATGACCTCTAAAGTAGCGCAAG	psc8
51	At5g67131	AJ412526	AATCAGGAAGATTCTGTCACCAA	GGAATGAGTGACAAAGCCAGA	Emil

TAB. R.1 – Séquences des différentes amorces définies à partir des séquences EST similaires aux gènes d'*Arabidopsis*.