Intérêt de marqueurs microsatellites pour l'étude de la diversité génétique des populations et variétés de luzerne pérenne

Bernadette Julier

Réf. **CB 47**

Contrat de branche 2005 / 2008

Rapport final (septembre 2008)

du programme de recherche :

« Intérêt de marqueurs microsatellites pour l'étude de la

diversité génétique des populations et variétés de luzerne pérenne »

Rappel des objectifs du projet

L'objectif de ce travail était de voir dans quelle mesure et avec quelle précision, les marqueurs microsatellites permettent de mesurer la diversité génétique et la structure génétique des populations et variétés de luzerne (*Medicago sativa*). Au cours de ce programme, les méthodes d'analyse correspondantes (génotypage, analyses statistiques) devaient être améliorées et transférées aux partenaires. L'analyse était basée sur un ensemble de variétés de luzerne représentatives du matériel cultivé en France (variétés flamandes et Provence) et sur un ensemble de populations naturelles de l'espèce, issues du pourtour méditerranéen et déjà connues pour de nombreux caractères (caractères phénotypiques et marqueurs cytoplasmiques). Pour les variétés, cette analyse moléculaire était couplée à une analyse phénotypique afin de déterminer dans quelle mesure la diversité révélée à l'aide de caractères phénotypiques reflète la diversité génétique sous jacente. De plus, une évaluation du déséquilibre de liaison entre marqueurs d'un chromosome devait être effectuée.

Le projet comprenait quatre aspects : (1) l'analyse de la diversité génétique dans et entre des variétés françaises de luzerne, (2) l'analyse de la diversité génétique entre les populations de luzerne, cultivées ou sauvages du pourtour méditerranéen, (3) le calcul du déséquilibre de liaison le long d'un chromosome, (4) la mise à disposition des outils du marquage moléculaire et des méthodes de calcul de la diversité entre populations chez des autotétraploïdes aux entreprises de sélection.

L'objectif de l'année 2007-2008 a principalement été de poursuivre l'analyse statistique des résultats de diversité génétique et de quantifier le déséquilibre de liaison à l'échelle d'un gène.

1. Diversité génétique de variétés françaises

1.1. Rappel des données disponibles et des résultats acquis

Sur 10 variétés (7 de type Flamande et 3 de type Provence), chacune représentée par 40 individus, une caractérisation phénotypique a été réalisée, en permettant une évaluation individuelle des plantes. Sur ces mêmes plantes, la diversité a été évaluée avec des marqueurs microsatellites supposés neutres.

Une large variabilité a été identifiée avec les caractères phénotypiques et avec les marqueurs. Les caractères phénotypiques permettent de distinguer trois groupes, Provence, Flamande, et Luzelle, variété qui forme un groupe à elle seule. Au sein du groupe Provence, les trois variétés sont bien différenciées, alors que dans le groupe Flamande, la variabilité intra-variétale recouvre la variabilité inter-variétale. Avec les marqueurs moléculaires, cette structuration n'est pas retrouvée. Néanmoins, on différencie clairement Luzelle des autres variétés. Avec ces marqueurs neutres, le F_{ST}, un indice qui mesure la part de variation expliquée par la structuration en variétés, n'est que de 0.013 (soit 1.3%). La variabilité intra-variétale est donc largement plus grande que la variabilité entre variétés.

Ces résultats ont été interprétés par un historique récent de l'introduction de la luzerne en Europe allié à une absence d'évolution génétique naturelle (pas de population spontanée), et de nombreux flux de gènes (flux de pollen voire de graines entre parcelles et échanges de semences entre agriculteurs). Tous ces facteurs s'opposent à la création d'une différenciation forte entre populations au niveau du fonds génétique, même si des caractères phénotypiques différencient les grands groupes de variétés.

1.2. Analyses statistiques supplémentaires

a/ Objectif

Nous avons cherché des explications sur la différence de structuration obtenue entre marqueurs moléculaires et données phénotypiques. Le F_{ST} calculé avec les données moléculaires est un paramètre synthétique qui peut être comparé à un Q_{ST} calculé avec des données phénotypiques. Quand Q_{ST} est supérieur à F_{ST}, ceci indique qu'il y a eu sélection sur le caractère phénotypique considéré. Quand les deux paramètres sont équivalents, le caractère phénotypique est assimilable à un caractère neutre, non soumis à la sélection.

b/ Calculs

Le Q_{ST} est calculé à partir des variances des effets variété et génotype :

$Q_{ST} = \delta^2_c/(2*\delta^2_g+\delta^2_c)$,

où δ^2_c est la variance de l'effet variété et δ^2_g la variance de l'effet génotype hiérarchisé à la variété.

L'écart-type de Q_{ST} a été calculé en utilisant une formule adaptée du calcul de l'écart type de la répétabilité (SE (r)) :

La répétabilité (r) a été calculée comme :

$r = \delta^2_c/(\delta^2_g+\delta^2_c)$,

D'où il suit que:

$Q_{ST} = (\delta^2_g+\delta^2_c)/(2*\delta^2_g+\delta^2_c)*r$

Et donc l'écart type de Q_{ST} est estimé comme :

$SE(Q_{ST}) = (\delta^2_g+\delta^2_c)/(2*\delta^2_g+\delta^2_c)*$ SE(r)

$= (\delta^2_g+\delta^2_c)/(2*\delta^2_g+\delta^2_c)*root(2*(1-r)^2*(1+(k-1)*r)^2/k*(k-1)*(C-1))$,

où k = nombre d'individus par variété = 40 et C = nombre de variétés = 10.

c/ Résultats

Les valeurs de Q_{ST} varient entre 0.00 et 0.39 selon les caractères et leur période de mesure (tableau 1). En moyenne, il est de 0.04 pour la date de floraison et atteint 0.19 pour le port. Pour la date de floraison, une seule mesure fournit un Q_{ST} significativement supérieur au F_{ST}. Pour la hauteur des tiges et la vitesse de croissance des tiges, un quart des valeurs sont supérieures au F_{ST} mais pour le port, toutes les valeurs sont significativement différentes du F_{ST}.

Tableau 1 - Estimation du Q_{ST} pour quatre caractères (hauteur des tiges MH, vitesse de croissance des tiges RG, port GH et date de floraison DF) mesurés dans deux lieux (Lusignan Lus and Connantre Con), pour deux cycles (c1/c2) et différentes années (ex 2003 = 03). Il est mentionné si l'intervalle de confiance de Q_{ST} inclus 0 et s'il inclut F_{ST}.

	Q_{ST}	Intervalle de confiance inclus 0	Intervalle de confiance inclus F_{ST}
MHc1Lus04	0.0359	x	x
MHc2Lus04	0.0652		x
RGc1Lus04	0.0603		x
RGc2Lus04	0.0534	x	x
GHc1Lus04	0.1438		
GHc2Lus04	0.1292		
DFc1Lus04	0.0026	x	x
DFc2Lus04	0.0205	x	x
MHc1Con04	0.0023	x	x
MHc2Con04	0.0890		x
RGc1Con04	0.0066	x	x
RGc2Con04	0.0066	x	x
MHc1Lus05	0.2048		
MHc2Lus05	0.1923		
RGc1Lus05	0.2557		
RGc2Lus05	0.1623		
GHc1Lus05	0.1652		
DFc1Lus05	0.0182	x	x
DFc2Lus05	0.0000	x	x
MHc1Con05	0.0038	x	x
MHc2Con05	0.0066	x	x
RGc1Con05	0.0974		
RGc2Con05	0.0094	x	x
GHc1Con05	0.3898		
DFc2Con05	0.0517	x	x
MHc1Lus06	0.0317	x	x
MHc2Lus06	0.1426		
RGc1Lus06	0.1107		
RGc2Lus06	0.0676		x
DFc1Lus06	0.0272	x	x
DFc2Lus06	0.0000	x	x
MHc1Con06	-	-	-
MHc2Con06	0.0290	x	x
RGc1Con06	0.0316	x	x
RGc2Con06	0.0426	x	x
MHc1Lus07	0.0397	x	x
MHc2Lus07	0.0389	x	x
RGc1Lus07	0.0812		x
RGc2Lus07	0.0766		x
GHc1Lus07	0.1291		
DFc1Lus07	0.0617		x
DFc2Lus07	0.1767		
Moyenne	0.0776		
Moyenne MH	0.0630		
Moyenne RG	0.0759		
Moyenne GH	0.1914		
Moyenne DF	0.0398		

d/ Interprétation

Lorsqu'on compare des caractères morphologiques et des marqueurs moléculaires, on suppose que les premiers ne sont pas neutres et que les seconds le sont. Quand les populations sont fortement structurées avec des caractères morphologiques mais pas avec des marqueurs moléculaires ($Q_{ST} > F_{ST}$), un rôle prédominant d'une sélection hétérogène entre populations est prédite. Au contraire, quand les populations sont différentiées autant pour les caractères morphologiques et pour les marqueurs neutres ($Q_{ST} = F_{ST}$), seule la dérive génétique explique la différenciation entre populations. Dans les populations naturelles, les valeurs de Q_{ST} sont généralement supérieures aux valeurs de F_{ST} suggérant que la sélection naturelle est une cause de la différentiation entre populations. Cette différentiation est prédite pour être plus forte pour les caractères à hérédité simple (mono ou oligogénique) que pour ceux à hérédité complexe (polygénique).

Dans notre étude, nous retrouvons une indication d'un effet de sélection sur les caractères morphologiques, les Q_{ST} sont majoritairement plus élevés que le F_{ST}. Cet effet de la sélection varie avec la complexité du caractère, puisque le port, un caractère connu pour avoir une hérédité simple (il est facilement sélectionné) présente le plus fort Q_{ST}. Chez les espèces cultivées, en dehors de la sélection naturelle, c'est la sélection artificielle qui génère la différenciation entre variétés. Cette sélection peut potentiellement aller dans le même sens que la sélection naturelle, les individus de meilleure valeur sélective pouvant être les individus générant la meilleure valeur agronomique (ex : des individus à longues tiges ont un avantage sélectif lié à l'aptitude à capter le rayonnement incident et sont aussi les plus productifs). Dans d'autres cas, sélections naturelle et artificielle vont dans des directions opposées. Dans notre étude, il y a une trace importante de sélection pour le port des plantes, critère important pour l'inscription des variétés. A l'opposé, la date de floraison qui n'est en général pas incluse comme critère de sélection chez la luzerne montre logiquement un faible Q_{ST}.

1.3. Estimation du nombre d'individus pour étudier la diversité des variétés de luzerne tétraploïde

La fiabilité des résultats d'une étude de diversité dépend du nombre de génotypes choisis pour représenter une population hétérogène. Dans la présente étude, nous avions choisi de représenter chaque variété par 40 individus. En parallèle, sur deux variétés (Mercedes et Symphonie), 120 individus ont été génotypés pour deux marqueurs microsatellites (MAA660456 et MTIC432). A partir des données, 20 sous-échantillons de 10, 20 et 40 individus ont été construits. Ils ont servi à calculer des F_{ST} par paire de sous-échantillons, de la même variété ou de variétés différentes.

Le nombre d'allèles révélés est beaucoup plus faible quand les sous-échantillons comportent 40 individus ou moins que quand ils comprennent 120 individus (Tableau 2). Ce nombre d'allèles décroît encore pour des sous-échantillons de 20 ou 10 individus. Ce sont des allèles rares qui sont perdus, les allèles fréquents sont révélés dans tous les cas.

Les valeurs F_{ST} pour des paires de sous-échantillons entre deux variétés (Tableau 3) étaient toujours significatives avec 40 individus, alors qu'elles n'étaient significatives que dans moins de la moitié des cas avec 10 individus. Des sous-échantillons de 20 individus donnent des résultats intermédiaires. Par ailleurs, quand on compare les sous-échantillons issus d'une même variété, ils sont toujours non différents, et cela quelque soit la taille de l'échantillon (Tableau 4).

Tableau 2: Pour deux marqueurs SSR (MAA660456 et MTIC432) et deux variétés (Mercedes et Symphonie), nombre d'allèles (moyenne, écart-type et coefficient de variation) pour 20 sous-échantillons de 40, 20 et 10 individus. Les mêmes données sont fournies sur l'échantillon de 120 génotypes.

	Mercedes (N=20)				Symphonie (N=20)		
	MAA660456	MTIC432	Total		MAA660456	MTIC432	Total
Sous-échantillons de 40 génotypes							
Nombre moyen	9.85	12.45	22.30		8.50	13.80	22.30
Ecart-type	1.27	1.23	1.87		0.51	1.40	1.75
CV	0.13	0.10	0.08		0.06	0.10	0.08
Sous-échantillons de 20 génotypes							
Nombre moyen	8.55	10.45	19.00		7.85	10.90	18.75
Ecart-type	1.10	1.32	1.59		0.67	1.92	2.07
CV	0.13	0.13	0.08		0.09	0.18	0.11
Sous-échantillons de 10 génotypes							
Nombre moyen	7.00	8.65	15.65		7.00	8.45	15.45
Ecart-type	0.97	1.79	2.13		0.79	1.50	1.93
CV	0.14	0.21	0.14		0.11	0.18	0.13
Total des 120 génotypes							
Nombre d'allèles	12	16	28		9	17	26

Tableau 3: Pour deux marqueurs et deux variétés, en prenant 20 sous-échantillons de 40, 20 et 10 individus, en considérant des paires de sous-échantillons de variétés différentes, pourcentage de F_{ST} significatifs (au seuil de P = 0.05), F_{ST} moyen et écart-type du F_{ST}.

	MAA660456	MTIC432	Deux marqueurs
Sous-échantillons de 40 génotypes			
F_{ST} significatifs	61.5	100.0	100.0
F_{ST} moyen	0.006	0.009	0.008
Ecart-type F_{ST}	0.006	0.008	0.004
Sous-échantillons de 20 génotypes			
F_{ST} significatifs	18.0	92.0	90.3
F_{ST} moyen	0.007	0.010	0.009
Ecart-type F_{ST}	0.011	0.012	0.007
Sous-échantillons de 10 génotypes			
F_{ST} significatifs	6.3	41.1	30.9
F_{ST} moyen	0.006	0.009	0.007
Ecart-type F_{ST}	0.021	0.016	0.013
Total 120 génotypes			
F_{ST}	0.006***	0.010***	0.008***

Tableau 4: Pour deux marqueurs et deux variétés, en prenant 20 sous-échantillons de 40, 20 et 10 individus, en considérant des paires de sous-échantillons de la même variété, pourcentage de F_{ST} significatifs (au seuil de P = 0.05), F_{ST} moyen et écart-type du F_{ST}.

	Intra Mercedes (N=190)			Intra Symphonie (N=190)		
	MAA660456	MTIC432	Total	MAA660456	MTIC432	Total
Sous-échantillons de 40 génotypes						
F_{ST} significatifs	0.0	0.0	0.0	0.0	0.0	0.0
F_{ST} moyen	-0.002	-0.001	-0.001	-0.003	-0.002	-0.002
Ecart-type F_{ST}	0.003	0.004	0.002	0.003	0.004	0.002
Sous-échantillons de 20 génotypes						
F_{ST} significatifs	0.5	0.0	1.1	0.5	2.6	0.5
F_{ST} moyen	-0.001	-0.002	-0.001	-0.002	-0.002	-0.002
Ecart-type F_{ST}	0.008	0.006	0.006	0.009	0.007	0.005
Sous-échantillons de 10 génotypes						
F_{ST} significatifs	3.7	3.2	3.2	0.0	0.5	0.0
F_{ST} moyen	-0.001	-0.004	-0.002	-0.005	-0.006	-0.005
Ecart-type F_{ST}	0.022	0.012	0.013	0.018	0.011	0.010

En conclusion, une taille d'échantillon de 40 individus par variété est raisonnable pour des études de diversité, dans lesquelles on cherche à comparer des variétés tétraploïdes relativement peu différentiées. Des échantillons de 10 individus sont trop petits pour obtenir des résultats fiables. Dans le cas de variétés assez différentiées, un échantillon de 20 individus par variété est certainement un bon compromis entre l'efficacité et le coût expérimental.

4. Estimation du déséquilibre de liaison chez la luzerne pour une étude d'association

Le déséquilibre de liaison (DL) décrit une association non aléatoire entre allèles à différents locus. Il peut être exploité pour réaliser des études de génétique, dites génétique d'association. Ainsi, si le DL est faible (la distance génétique à partir de laquelle on n'observe pas de déséquilibre est petite), une association entre le polymorphisme allélique à un gène et le polymorphisme phénotypique démontrera l'implication du gène dans le caractère (« approche gène candidat »). Au contraire, si le DL est fort, on déduira d'une association entre un marqueur et un caractère que la région du marqueur est impliquée dans la variation du caractère (approche « genom scan »).

Chez les espèces allogames, le DL est généralement faible, c'est donc le résultat attendu chez la luzerne.

Partant de ce résultat, nous avons cherché à évaluer le DL dans une séquence de l'ordre du millier de paires de bases. Nous avons choisi un gène qui pourrait expliquer chez la légumineuse modèle *Medicago truncatula* les variations pour la date de floraison et l'allongement des tiges (thèse de J.B. Pierre, INRA Lusignan, J.B. Pierre et al, TAG 2008). Ce gène, Constans, est impliqué dans le déterminisme de la date de floraison d'autres espèces. Il comprend quatre introns et trois exons (figure 1).

4.1. Méthodologie

Le choix a été fait de chercher à amplifier deux portions du gène Constans, au début et à la fin du gène. Une première étape a été de définir des amorces utilisables sur la luzerne et amplifiant spécifiquement ce gène, en partant des données de séquences disponibles sur *M. truncatula*.

Ensuite, pour estimer le DL, il est important d'avoir la séquence des quatre allèles du gène pour chaque individu. On a choisi de travailler sur une seule variété, Mercedes, représentée par 40 individus, et sur la fin du gène. Les amorces ont permis d'amplifier la portion du gène par PCR, et les fragments ont été clonés dans des vecteurs bactériens. Pour chaque individu, huit clones ont été obtenus, et la taille des fragments clonés a été déterminée après PCR et migration sur gel d'agarose. Les fragments ont été séquencés par une société de service.

Dans l'objectif d'une étude d'association, les deux portions du gène ont été séquencées sur les 400 individus (10 variétés x 40 individus), après PCR avec les amorces spécifiques. Le séquençage a été réalisé directement sur les produits PCR par les deux extrémités, le polymorphisme SNP étant visible sur les résultats de séquençage.

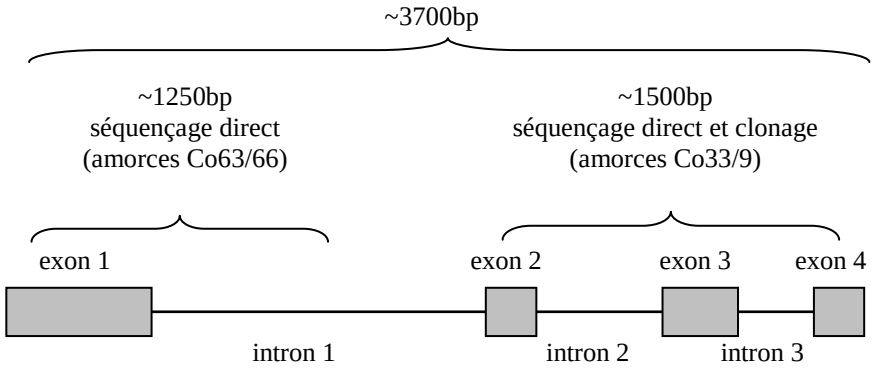
4.2. Résultats

a/ Définition d'amorces

Sur la fin du gène, une paire d'amorces utilisée sur *M. truncatula* (Co33/Co9) permet d'amplifier deux bandes chez la luzerne. L'une des bandes a la taille attendue d'environ 1500 pb, alors que la seconde bande est d'environ 1700 pb. Un séquençage direct de ces deux bandes pour deux individus dans les deux sens montre que ces fragments correspondent bien au gène Constans de *M. truncatula*. Au milieu de la séquence, il y a probablement un grand insert non décrit chez *M. truncatula* et qui génère la bande la plus longue chez la luzerne. Cette paire d'amorces a été utilisée pour cloner cette portion du gène chez 40 individus de Mercedes et estimer le LD, et pour le séquençage direct des 400 individus.

Pour le début du gène, il n'y avait pas d'amorces disponibles chez *M. truncatula*. Des nouvelles amorces ont donc été définies dans les exons 1 et 2, et testées sur la luzerne. Un couple d'amorces (Co53/Co61) donne une amplification faible mais nette d'une bande d'environ 2000 pb. Pour deux individus, cette bande a été séquencée directement, et l'alignement avec la séquence de *M. truncatula* a montré que le fragment amplifié correspondait bien au gène Constans. Cependant le faible niveau d'amplification et la taille élevée du fragment étaient un problème pour la suite des opérations. Pour cette raison, de nouvelles amorces ont été définies à partir des séquences obtenues sur la luzerne. Le couple d'amorces Co63/Co66 amplifie clairement deux fragments d'environ 1250 et 1600 pb. Ce couple d'amorces a été utilisé pour le séquençage direct des 400 individus.

Figure 1:Schéma du gène Constans chez *Medicago truncatula.* Les portions du gène amplifiées pour clonage ou séquençage direct chez la luzerne sont indiquées. Les longueurs mentionnées correspondent à celles de *M. truncatula*



b/ Estimation du déséquilibre de liaison

Pour calculer le DL, il faut connaître la phase entre les différentes mutations (ou SNP pour Single Nucleotide Polymorphism) : est-ce que les SNP sont portés par le même chromosome ou sont sur des chromosomes homologues. Pour cela, la séquence doit être obtenue après isolement des fragments PCR, ce qui est réalisé par leur clonage dans un vecteur bactérien. Cette opération de clonage demande beaucoup de travail et est onéreuse. Elle n'a été réalisée que sur Mercedes, représentée par 40 individus.

Des PCR ont été réalisées avec une TAQ haute fidélité, en utilisant les amorces Co33/Co9 et l'ADN génomique des 40 individus de Mercedes. Les produits PCR ont été purifiés, ligués dans un plasmide qui a servi à transformer des bactéries *E. coli*. Pour chaque individu de luzerne, huit colonies transformées ont été isolées et une PCR a été réalisée sur chaque colonie (8 colonies x 40 individus = 320 réactions PCR). Après migration sur gel d'agarose, on constate que 207 colonies fournissent la bande légère (1500 pb), 59 colonies ont la bande lourde (1700 pb) et 54 n'ont donné aucune amplification. Au bilan, 20 individus portaient les deux bandes, quatre individus n'avaient que la bande légère et 15 individus que la bande lourde. Le produit PCR a été séquencé dans les deux directions. Quand deux bandes différentes ont été détectées pour un individu, elles ont toutes deux été séquencées. Pour seulement un individu, il n'y avait pas de séquence utilisable. Au total, 59 séquences ont été obtenues.

Tableau 5: SNP (A) et insertions/délétions (B) détectées dans une portion du gène Constans chez la luzerne

A) Position du SNP (pb)	Nombre d'individus avec une base différente
16	**11**
73	**10**
114	5
205	4
332	**27**
354	2
446	**9**
504	5
872	2
923	**16**
937	**6**
975	**11**
1170	**11**
1344	2
1371	4
1701	**20**
1740	**6**
1829	5
1830	5
1935	**26**
1948	**7**

B) Position de l'insertion/délétion (bp)	Nombre d'individus avec insertion
559-785	14 (cf. fig. 2)
873	54
977-1163	11 (cf. fig. 2)
1268-1275 (1268-1313)	54 (57)
1865-1868	2

En gras, les SNP présents à une fréquence supérieure à 10%.

Au total, 82 SNP ont été identifiés, dont 61 singletons (présents dans une seule séquence). Sur les 21 autres SNP, seulement 12 sont présents dans plus de 10% des séquences, ce sont ces 12 SNP qui sont vraiment informatifs (Tableau 5). De plus, sept insertions/délétions ont été détectées. Deux sont des singletons, trois autres sont de moins de 50 pb (et ne peuvent pas expliquer la différence de taille des deux bandes) et présentent dans moins de 10% des individus. Enfin, trois inserts pourraient expliquer la bande lourde (1700 pb) (Figure 2) :

insert a: pour 10 séquences, un gros insert de 227 pb après 559 pb dans la séquence

insert b: pour quatre séquences, un insert de même taille au même endroit, mais en sens inverse

insert c: pour 10 (+1) séquences, un gros insert de 187 pb après 977 pb dans la séquence. Cette portion correspond à la jonction du séquençage de chaque sens, elle est relativement incertaine. La présence de l'insert est sûre, sa longueur l'est moins. Pour un individu, il n'y a que le début et la fin de l'insert, et comme la séquence est de bonne qualité, on peut voir que l'insert n'est que de 61 pb.

Au début ou à la fin des inserts a et b, nous avons trouvé cinq différents cas de délétions mineures (figure 2). Ces inserts à 559 pb et ces petites délétions font penser à un transposon qui pourrait s'intégrer dans un sens ou dans un autre, et laisser des « traces » de son passage (les petites délétions) lorsqu'il s'insère ou s'excise.

Figure 2: Schéma des principaux inserts dans une partie du gène Constans chez la luzerne

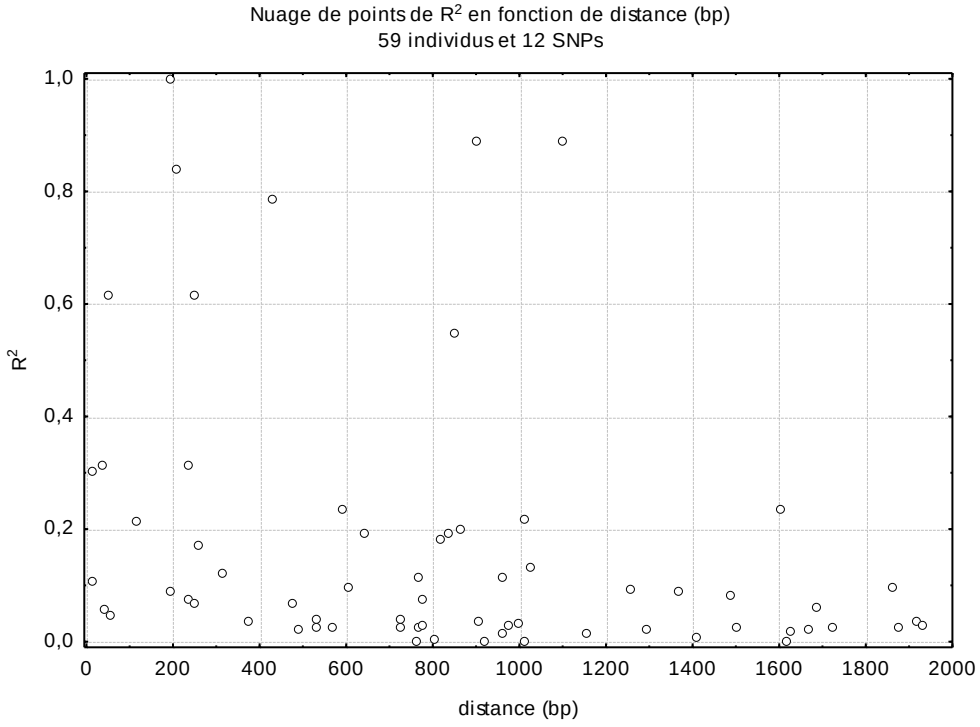
```
……— 559 ———……———— 758 ————…… 977 ———……——— 1163……bp

CTATTA  insert a (227bp) AGTAATATACTAG……TG-délétion 187bp---GG    10 individus
CTATTA  insert b (227bp) AGTAATATACTAG……TG-délétion 187bp---GG     4 individus

CTATTA--délétion 232bp-------ATATGCTAG……AG insert c (187bp) GG    11 individus

CTATTAC-délétion 227bp---AGTAATATACTAG……TG-délétion 187bp---GG     5 individus
CTATTAC-délétion 229bp-----TAATATACTAG……TG-délétion 187bp---GG    10 individus
CTATTA--délétion 232bp-------ATATACTAG……TG-délétion 187bp---GG    17 individus
CT------délétion 245bp----------------  TG-délétion 187bp---GG     2 individus
```

Figure 3: Evolution du DL en fonction de la distance dans le gène Constans chez la luzerne



Le coefficient R^2 (qui représente le DL) a été calculé avec les 12 SNP (un total de 66 paires). Seulement huit paires de SNP ont un R^2 supérieur à 0.4, parmi elles, cinq sont dans des distances inférieures à 400 pb et trois entre 800 et 1100 pb (figure 3). Pour six de ces huit paires de SNP, les quatre SNP suivants sont retrouvés :

1. 1 pb avant l'insert c (975 pb) avec une base différente dans les 11 séquences ayant l'insert c
2. 7 pb après la fin de l'insert c (1170 pb) avec une base différente dans les 11 séquences ayant l'insert c
3. 54 pb avant l'insert c (923 pb), avec une base différente dans les 11 séquences ayant l'insert c, les quatre séquences avec l'insert b et une autre séquence

4. 904 pb avant l'insert c (73 pb) avec une base différente dans 10 des 11 séquences ayant l'insert c (ce SNP est impliqué dans les trois paires de SNP significatifs avec une distance entre SNP de 800 à 1100 pb).

Finalement, ces SNP générant un R^2 fort sont généralement liés à l'insert c. Le cas du SNP à 73 pb est particulièrement intéressant puisqu'il montre presque 100% de relation avec l'insert c alors qu'il en est à 900 pb. De plus, ce SNP à 73 pb est dans l'exon 2. La conservation de l'haplotype avec l'insert c et le SNP à 73 pb pourrait être important pour le devenir des plantes (survie ou performance) et pourrait donc être associé à un caractère.

En résumé, une très faible distance de DL a été trouvée dans cette partie du gène Constans, comme il était attendu chez la luzerne, espèce fortement allogame. Si une association est détectée entre un SNP de ce gène et un caractère, cela signifiera que la mutation générant la variation phénotypique est très proche du SNP. L'approche « gène candidat » pour les études d'association est bien celle qui sera efficace.

4.3. Perspectives : Etude d'association

Le séquençage direct des 400 individus pour les deux portions du gène Constans permettra de fournir des données pour une étude d'association. Le séquençage aura lieu dans une seule direction, ce qui fournira environ 500 pb dans chacun des fragments. La détection de SNP dans ces fragments sera mise en relation avec les données phénotypiques disponibles.

5. Valorisation

Deux articles sont en cours de préparation et seront prochainement soumis à des revues internationales :

Herrmann D., Flajoulot S., Barre P., Huyghe C., Ronfort J., Julier B. Comparison of morphological traits and SSR to analyse diversity and structure of alfalfa cultivars. BMC Genetics, en préparation.

Herrmann D., Flajoulot S., Barre P., Huyghe C., Ronfort J., Julier B. Optimal sample size for diversity studies based on codominantly coded SSR markers in tetraploid populations Plant Breeding (communication courte) en préparation

Une communication par poster a été produite à l'occasion du congrès nord-américain sur l'amélioration de la luzerne (NAAIC) qui se tenait en juin 2008 à Dallas (participation de B. Julier, S. Flajoulot et D. Herrmann) :

Herrmann D., Flajoulot S., Barre P., Huyghe C., Ronfort J., Julier B. (2008) Comparison of morphological traits and SSR markers to analyze genetic diversity of alfalfa cultivars. NAAIC, http://www.naaic.org/Meetings/National/2008meeting/proceedings/Herrman.pdf