Le séquençage des génomes de plantes : les acquis

Michel Delseny

Laboratoire Génome et Développement des Plantes (LGDP)
UMR 5096
CNRS, IRD
Université de Perpignan
66860 Perpignan
France
<delseny@univ–perp.fr>

Résumé

Cette revue présente un historique du séquençage des génomes de plantes, ses principaux résultats et les difficultés rencontrées. Actuellement, seuls les génomes d'*Arabidopsis* et du riz ont été séquencés et assemblés avec un taux d'erreur minimal, mais plusieurs brouillons détaillés sont d'ores et déjà disponibles pour une dizaine d'espèces et le séquençage d'une bonne vingtaine d'autres est en cours. Ces travaux permettent de dresser des catalogues des gènes de chaque plante. Ils permettent aussi d'analyser l'évolution des génomes. L'une des grandes surprises révélées par le séquençage est l'existence, au cours de l'évolution, d'évènements de duplications chromosomiques globales (WGD) accompagnés du maintien, de la perte ou de la divergence des deux copies d'un même gène.

Mots clés : génomique végétale ; séquençage.

Thèmes : métabolisme ; pathologie ; productions végétales.

Summary

Sequencing plant genomes: accomplishments

This review summarizes our progress in sequencing plant genomes. So far, the genomes of only two plants, Arabidopsis and rice, have been sequenced and assembled accurately. Detailed drafts have however been obtained for a dozen species and sequencing of about twenty other species is in progress. The main result of such experiments is to provide an exhaustive catalogue of plant genes while giving insight into their individual evolution as well as the global evolution of plant genomes. A major surprise revealed by sequencing *Arabidopsis* was the observation that several rounds of whole genome duplications (WGD) occurred during the plant's evolution. After duplication, duplicated copies are either lost, conserved or diverge so that their functions differentiate

Key words: plant genomics; sequencing.

Subjects: metabolism; pathology; vegetal productions.

L'achèvement du séquençage d'un génome de la plante modèle *Arabidopsis* en 2000, puis de celui d'un génome de riz en 2004, constitue des étapes clés dans notre connaissance du génome des plantes. Ces deux percées ont permis un essor sans précédent, de la biologie végétale. En effet, un catalogue à peu près complet de l'ensemble des gènes d'un génotype d'une espèce de plante est disponible, même si l'on est encore loin d'avoir identifié la fonction de chacun d'eux. Des outils performants d'analyse de l'expression des gènes ont été développés avec la mise au point de puces à ADN et la réalisation de bases de données transcriptomiques et protéomiques. Des collections de mutants ont été réalisées, permettant d'associer un gène à un phénotype donné et, le plus souvent, d'isoler le gène muté, soit par étiquetage avec un T-DNA[1] ou un transposon, soit par clonage positionnel. L'ensemble de ces technologies a conduit à élucider, au moins chez *Arabidopsis*, au niveau moléculaire, la plupart des grandes fonctions d'une plante en termes de voies métaboliques et de voies de signalisation.

Dans cette présentation, nous retraçons l'histoire du séquençage des génomes de plantes, évoquons les problèmes rencontrés et résumons ce que nous avons appris de leur organisation et de leur évolution.

[1] Portion de plasmide d'*Agrobacterium* transférée au génome de la plante.

doi: 10.1684/agr.2009.0343

Tirés à part : M. Delseny

Un bref historique du séquençage du génome des plantes

Des progrès technologiques

Le séquençage des génomes est associé à toute une série de progrès techniques cruciaux. Dans la deuxième moitié des années 1970, le clonage de gènes et les méthodologies de base du séquençage sont mis au point, mais les premiers gènes de plantes ne sont caractérisés qu'au début des années 1980. Au cours des 10 années qui suivent, les méthodes se perfectionnent et de nombreux gènes sont isolés et caractérisés, mais le débit reste artisanal : quelques ADNc[2] ou et des gènes codant pour des protéines abondantes sont caractérisés, mais, à quelques exceptions près, il est très difficile d'isoler un gène repéré par une mutation. De plus, les fragments de génome isolés demeurent de petite taille. Le début des années 1990 est marqué par une série de nouveaux progrès. L'utilisation de traceurs radioactifs pour le séquençage est remplacée par des marqueurs fluorescents ce qui permet la commercialisation des séquenceurs automatiques de première génération. Des nouveaux vecteurs pour fragments de grande taille, les YAC[3] et les BAC[4] sont mis au point, principalement sur *Arabidopsis*, et permettent le développement de cartes physiques[5]. Des programmes ambitieux, comme le séquençage du génome de la levure, du vers nématode *Caenorhabditis elegans*, de la drosophile ou d'*Arabidopsis thaliana* se mettent en place pour préparer celui du génome humain. Enfin, une stratégie nouvelle de séquençage est proposée par Craig Venter (Adams *et al.*, 1992) : le séquençage des EST[6] consiste à séquencer de façon aléatoire et partielle des collections d'ADNc ; on peut ainsi dresser, à faible coût, un catalogue assez complet des gènes exprimés à un moment donné ou dans un tissu donné.

[2] copie ADN d'un ARN messager ou ADN complémentaire.
[3] *Yeast artificial chromosome.*
[4] *Bacterial artificial chromosome.*
[5] Organisation et mise en continuité des différents BAC pour chaque chromosome par l'analyse des séquences communes de l'ensemble des BAC.
[6] Séquences partielles de gènes exprimés (*Expressed Sequenced Tag*).

Programmes d'EST

Chez les plantes, trois projets d'EST ont démarré à peu près simultanément en 1992-1993 : en France et aux États-Unis, sur *Arabidopsis* (Hofte *et al.*, 1993, Newman *et al.*, 1994, Cooke *et al.*, 1996) et, au Japon, sur le riz (Sasaki *et al.*, 1994). Alors qu'à cette époque, moins de 200 gènes avaient été partiellement caractérisés toutes plantes confondues, en quelques années ce chiffre est passé à plusieurs milliers sur ces deux espèces. Ces premiers travaux ont marqué le pas lorsque le séquençage génomique a commencé, mais il a été relativement vite réalisé que des collections de séquences d'ADNc – si possible de pleine longueur – étaient indispensables à l'annotation des séquences génomiques. De ce fait, cette stratégie s'est généralisée à la majorité des plantes cultivées, si bien qu'aujourd'hui, plusieurs millions d'EST de plantes sont disponibles *(tableau 1)*. Pour une vingtaine d'espèces, plus de 200 000 EST sont disponibles et, pour une trentaine d'autres, entre 50 000 et 200 000 le sont également.

Séquençage des génomes modèles

Le séquençage génomique a commencé dès 1993, par un projet pilote européen sur le chromosome 4 d'*Arabidopsis* (Bevan *et al.*,1998), puis s'est accéléré sous l'impulsion d'un consortium international, aboutissant à une séquence quasi complète en 2000 (The Arabidopsis Genome Initiative, 2000). Dès 1997, il était clair que le séquençage d'un génome complet de plante était possible et que celui d'*Arabidopsis* serait achevé plus tôt

Tableau 1. Évolution récente du nombre d'EST de plantes dans la base dbEST.

Table 1. Recent devlopments in the number of plant ESTs in the ESTdb.

dbEST release	**19/12/98**	**25/01/05**	**13/04/07**	**10/10/08**
Entrées totales		***25 091 510***	***42 964 657***	***57 206 420***
Homo sapiens		6 009 065	7 974 440	8 138 094
Arabidopsis thaliana	3 7628	326 202	1 276 131	1 526 133
Zea mays	2752	417 056	1 161 241	1 464 859
Glycine max		344 562	371 817	1 317 957
Oryza sativa	35192	298 857	1 211 154	1 220 876
Triticum aestivum		587 650	1 050 131	1 051 300
Brassica napus	1 434	45 906		596 249
Hordeum vulgare		375 187	437 728	478 734
Panicum virgatum				436 535
Phaseolus coccineus		20 120		391 138
Vitis vinifera		147 300	320 503	352 984
Pinus taeda	1 953	173 680		328 628
Picea glauca		55 108		272 464
Gossypium hirsutum	1 327	23 899		268 565
Solanum lycopersicum		153 911	199 875	258 830
Malus x domestica		183 732		255 659
Medicago truncatula		216 645	225 522	250 471
Saccharum officinalis		246 301	246 301	246 373
Nicotiana tabacum		27 014		240 440
Solanum tuberosum		192 038	219 765	231 116
Sorghum bicolor		208 196	208 466	209 814
Citrus sinensis		75 579	94 704	203 752
5 espèces				> 100 000
25 espèces				> 50 000

La première ligne indique les dates auxquelles le nombre d'EST pour chaque espèce est déterminé. À titre de comparaison, le nombre total d'EST dans la base de données et le nombre d'EST humaines sont indiquées dans les deux lignes suivantes. Seules ont été retenues les espèces de végétaux supérieurs pour lesquelles le nombre d'EST est actuellement supérieur à 200 000. Source : base dbEST (http://www.ncbi.nlm.nih.gov/dbEST).

que prévu. Il était également déjà clair que la connaissance du génome d'*Arabidopsis* serait insuffisante pour comprendre le génome des espèces cultivées. Le riz paraissait le meilleur candidat, du fait des ressources disponibles, de la taille de son génome et des évidences de colinéarité[7] avec les génomes des autres céréales. Un consortium international s'est alors constitué, comme pour *Arabidopsis*. L'utilisation de séquenceurs automatiques de nouvelle génération, la concentration des activités dans quelques centres performants, ainsi qu'une compétition entre consortium public et firmes privées ont permis d'atteindre l'objectif dès 2004 : la publication d'un premier brouillon (Goff *et al.*, 2002 ; Yu *et al.*, 2002 ; International Rice Sequencing Project, 2005).

Séquençage des autres génomes

Dès 2000, il est devenu réaliste de séquencer, plus ou moins complètement, les génomes de quelques autres espèces cultivées dans des délais rapprochés. Diverses stratégies sont employées : des stratégies BAC à BAC, dans lesquelles une carte physique sert de guide et permet un séquençage ordonné et un assemblage complet et précis, ou bien des stratégies globales (« *shotgun* »), dans lesquelles le génome est débité en fragments aléatoires de différentes tailles, séquencés en aveugle à partir de chaque extrémité et que l'on assemble ensuite à l'aide de logiciels dédiés. Cette dernière stratégie, beaucoup plus rapide et moins chère, conduit à une ébauche du génome et laisse de nombreux trous et erreurs d'assemblage dans la séquence, mais permet de repérer facilement les gènes d'intérêt. Selon les espèces, l'une ou l'autre stratégie, ou une combinaison des deux, sont utilisées. Cependant, pour ensuite utiliser la séquence dans des programmes d'amélioration des plantes, il est indispensable de l'ancrer sur la carte génétique. Les trois dernières années ont vu la publication des ébauches détaillées des génomes du peuplier (Tuskan *et al.*, 2006), de la vigne (Jaillon *et al.*, 2007 ; Velasco *et al.*, 2007), du papayer (Ming *et al.*, 2008) et du sorgho (Paterson *et al.*, 2009). Celles de *Brachypodium*, du maïs, du ricin et du lotier sont disponibles dans les bases de données.

[7] Arrangements identiques des séquences sur certaines portions de chromosomes.

Analyse du génome d'*Arabidopsis* et du riz

Annotation physique et fonctionnelle

L'analyse explicative du génome s'est révélée beaucoup plus complexe que prévue. Après séquençage, l'objectif est d'identifier les gènes, de déterminer leur nombre, leurs limites exactes et ensuite leur fonction. Cette analyse constitue l'annotation. Malgré la grande qualité des séquences (moins d'une erreur pour 40 000 nucléotides), les premiers logiciels utilisés, dérivés de l'analyse des génomes bactériens ou de levure, ne prédisent pas correctement les limites des gènes, en particulier les jonctions exon/introns, ne font pas toujours la différence entre des exons appartenant à deux gènes adjacents ou des exons d'un même gène, ratent des phases de lecture… Enfin, au moment où la séquence complète est rendue publique, un certain nombre de propriétés du génome sont encore inconnues. C'est le cas des gènes ne codant pas pour des protéines, mais déterminant des petits ARN, tels ceux des petits ARN nucléolaires (snoRNA), des micro-RNA (miRNA), ou des *silencing RNA* (siRNA), de l'existence de jonctions exon/intron non canoniques, ou encore d'un gradient 5' vers 3' de teneur en GC[8] de certains gènes de céréales. La découverte de ces nouvelles propriétés a permis d'améliorer les logiciels d'analyse. Mais la validation la plus sûre d'une prédiction reste l'expérimentation. Ainsi, chez *Arabidopsis*, le nombre de gènes est passé de 26 000 à près de 30 000 (Ilic *et al.*, 2007) répartis en environ 11 600 familles. Cela n'a pu se faire que parce que la séquence génomique a été systématiquement confrontée à celles des EST et des ADNc complets et que ces collections de séquences exprimées sont de plus en plus complètes. Une autre complication vient du fait que nombre d'ADNc correspondent à des ARNm[9] qui ont subi un épissage alternatif et que certains gènes sont en fait des pseudogènes. Un problème supplémentaire est apparu avec la séquence du riz, beaucoup plus riche en séquences répétées telles que les rétrotransposons[10].

[8] Paire de bases Guanine Cytosine.
[9] ARN messagers
[10] Éléments génétiques capables de se déplacer dans le génome.

Ces éléments contiennent des séquences reconnues comme des gènes par les logiciels de prédiction. De ce fait, les premières annotations automatiques du génome du riz ont surévalué le nombre de gènes en comptabilisant des séquences associées aux rétroéléments. L'annotation des gènes déterminant les différentes classes de petits ARN a rapidement progressé depuis leur découverte avec l'amélioration des logiciels de prédiction et la mise en œuvre de méthodes de séquençage à haut débit (Ben Amor *et al.*, 2009, Kasschau *et al.*, 2007, Zhou *et al.*, 2009). Au-delà de l'annotation physique, c'est l'annotation fonctionnelle qui intéresse le biologiste : lors de la publication de la séquence d'*Arabidopsis*, en 2000, moins de 10 % des gènes avaient une fonction prédite ayant reçu un début de validation expérimentale. L'objectif, depuis cette date, a donc été de proposer une fonction pour chacun des 30 000 gènes identifiés. Si les progrès sont tangibles, avec des résultats expérimentaux pour environ 40 % des gènes et des prédictions fiables pour près de 80 %, il reste encore 20 % des gènes pour lesquels nous n'avons aucune idée de ce pour quoi ils codent. Les données sont moins satisfaisantes pour le riz.

Les principaux outils d'annotation

La première stratégie est l'annotation automatique, à l'aide de logiciels prédictifs de la structure des gènes, intégrés sur des plateformes d'annotation qui permettent de naviguer dans les bases de données et d'utiliser tous les éléments disponibles. L'annotation est ainsi un processus dynamique, qui n'est pratiquement jamais terminé. Deux approches complémentaires permettent d'améliorer l'annotation physique. La première consiste à isoler et caractériser des ADNc pleine longueur. L'observation d'une séquence transcrite prouve que la région correspondante de l'ADN est exprimée, au moins au niveau ARN. L'analyse des ADNc pleine longueur d'*Arabidopsis* ou de riz a permis de découvrir plusieurs centaines de gènes que les logiciels de prédiction n'avaient pas réussi à identifier (Seki *et al.*, 2002). Cette stratégie continue de se développer avec l'utilisation des puces chevauchantes (*Tiling arrays*) (Gregory *et al.*, 2008) et les analyses du transcriptome par séquençage à haut débit (Lee *et al.*, 2005). Pour aller au-delà, il faut montrer qu'une protéine

correspondante à un ARNm existe réellement : c'est tout l'apport de la protéomique à l'analyse des génomes. De même, les approches protéomiques ont révélé l'existence de centaines de gènes non détectés par les logiciels de prédiction ou l'analyse des ADNc (Baerenfaller *et al.*, 2008). Une deuxième approche est la génomique comparative, qui repose sur l'existence d'un répertoire de gènes en grande partie partagé entre tous les êtres vivants et dont la structure est assez largement conservée. Lorsque des séquences présentent un pourcentage d'identité significatif, on dit qu'elles sont homologues. La détection dans le génome d'une espèce d'une séquence homologue à un gène identifié chez une autre espèce constitue un indice fort, mais pas suffisant, que les deux gènes ont une fonction similaire dans les deux espèces (Ayele *et al.*, 2005, Cooke *et al.*, 2007). Cette approche, très limitée au début, devient maintenant praticable à grande échelle en raison de l'accumulation exponentielle de séquences génomiques et d'EST. Elle a aussi nécessité la mise en place de définitions fonctionnelles des classes de gènes utilisées par l'ensemble de la communauté scientifique (Berardini *et al.*, 2004).

L'annotation fonctionnelle reste très empirique. Le premier outil est la recherche d'homologies à l'aide des logiciels d'alignement, mais une validation expérimentale reste indispensable. Une deuxième approche repose sur l'existence ou la création de mutants que l'on peut complémenter. On isole le gène correspondant par clonage positionnel ou étiquetage par un T-DNA ou un transposon. La complémentation de la mutation en réintroduisant le gène établit que celui-ci a un rôle dans le déterminisme du phénotype. Il existe de multiples variantes de cette approche : complémentation de mutants de bactéries, de levures, de plantes ou de cellules animales. La réalisation de grandes collections de mutants, tant chez *Arabidopsis* (Alonzo *et al.*, 2003) que chez le riz (Sallaud *et al.*, 2004), a permis de saturer quasiment ces deux génomes avec des insertions dont les séquences flanquantes ont été déterminées. Cet outil permet d'obtenir des mutants pour à peu près n'importe quel gène : il suffit alors d'interroger les bases de données et de commander les lignées correspondantes. La limite de cette approche est que tous les mutants n'ont pas nécessairement un phénotype évident et il est souvent nécessaire de construire des mutants multiples pour établir une fonction. Cela est souvent dû au fait que les gènes appartiennent souvent à des familles multigéniques et que leurs fonctions sont en partie redondantes.

Lorsque ces deux types d'approches ont échoué, il reste la possibilité de comparer le profil d'expression du gène inconnu à celui de l'ensemble des autres gènes et à déterminer s'il rentre dans l'une des catégories de profil d'expression. On peut ainsi souvent associer un gène à un processus connu et mieux cerner sa fonction. Cette approche est devenue possible grâce à la mise en place de bases de données normalisées dans lesquelles sont archivées les expériences de puces à ADN (Zimmermann *et al.*, 2004). On peut aussi rechercher quels sont les partenaires d'une protéine inconnue. On réussit ainsi parfois à associer un gène et sa protéine à un réseau d'interactions qui renseigne sur sa fonction. Malgré tous ces efforts, la difficulté majeure de la caractérisation fonctionnelle reste la redondance d'un grand nombre de gènes et la pauvreté de notre capacité à analyser les phénotypes.

Duplications du génome d'*Arabidopsis*

Dès l'analyse des EST, l'existence généralisée de familles multigéniques a été révélée. Il y a ainsi trois à quatre gènes pour chacune des quelque 80 protéines ribosomiques d'*Arabidopsis* (Barakat *et al.*, 2001) et nous avons cherché à en comprendre la logique d'organisation. Ceci nous a conduits à mettre en évidence que de grandes régions du génome d'*Arabidopsis* avaient apparemment été dupliquées au cours de l'évolution (Blanc *et al.*, 2000). Cette observation a donné lieu à toute une série de travaux qui ont conduit à une nouvelle vision de la façon dont les génomes de plantes ont évolué.

Une première interrogation a porté sur la date de ces duplications. Les premiers résultats ont été particulièrement confus, car en partie biaisés par la mauvaise qualité de l'annotation initiale. Désormais on s'accorde pour admettre que le génome d'*Arabidopsis* a subi au moins deux duplications globales de son génome (WGD, *Whole Genome Duplication*), l'une relativement récente il y a à peu près 20 millions d'années, l'autre plus ancienne, probablement vers 50-70 millions d'années (Blanc *et al.*, 2003). Ainsi, la conclusion qui s'impose est que le génome d'*Arabidopsis*, choisi initialement en raison de sa simplicité apparente, est en fait un *patchwork* de segments génomiques dupliqués à différentes périodes de l'évolution. Le génome se décompose en une centaine de blocs dupliqués. Dans les blocs les mieux conservés, 30 % des gènes sont conservés par paires ; ce pourcentage décroît à 15-16 % dans les blocs plus anciens. Les autres gènes correspondent soit à des paires dont l'une des copies a été perdue, soit, plus rarement, à des gènes insérés après duplication. On peut alors proposer un schéma d'évolution du génome par duplications successives suivies d'une phase de perte et d'élimination de certains des gènes dupliqués *(figure 1)*.

L'existence de régions dupliquées et de copies paralogues de certains gènes pose la question de leur signification biologique. Deux copies peuvent avoir différents devenirs : l'élimination pure et simple de l'une des deux copies, le maintien de deux copies complètement ou partiellement redondantes ou bien ayant acquis une spécialisation fonctionnelle. Les gènes paralogues sont des gènes homologues qui ont divergé après une duplication au sein d'une même espèce. On les oppose aux gènes orthologues qui sont des gènes appartenant à des espèces différentes et qui ont divergé à partir d'un gène unique d'une espèce ancestrale). Le jeu des duplications et de leur évolution conduit donc à une augmentation de la diversité fonctionnelle des gènes et permet ainsi sans doute aux plantes de survivre à un environnement climatique et biotique en constant changement.

Indépendamment de ces WGD, le génome d'*Arabidopsis* contient environ 16 % de gènes dupliqués en tandem. Leur mode d'évolution est clairement distinct des précédents, faisant appel à des processus de recombinaison locale et de *crossing-over* inégaux, qui ont souvent pour effet de profondément modifier la structure des promoteurs. On va donc trouver parmi ces gènes paralogues davantage de cas de subfonctionnalisation et de spécialisation au niveau de l'expression que parmi les paralogues résultant de WGD. La classification des gènes en grands groupes ontologiques révèle que le devenir des gènes paralogues, après duplication, n'est pas aléatoire (Blanc et Wolfe, 2004b). Certains gènes, comme ceux des facteurs de transcription, sont préférentiellement conservés en paires qui se spécialisent et acquièrent de nouvelles fonctions. À l'opposé, les gènes

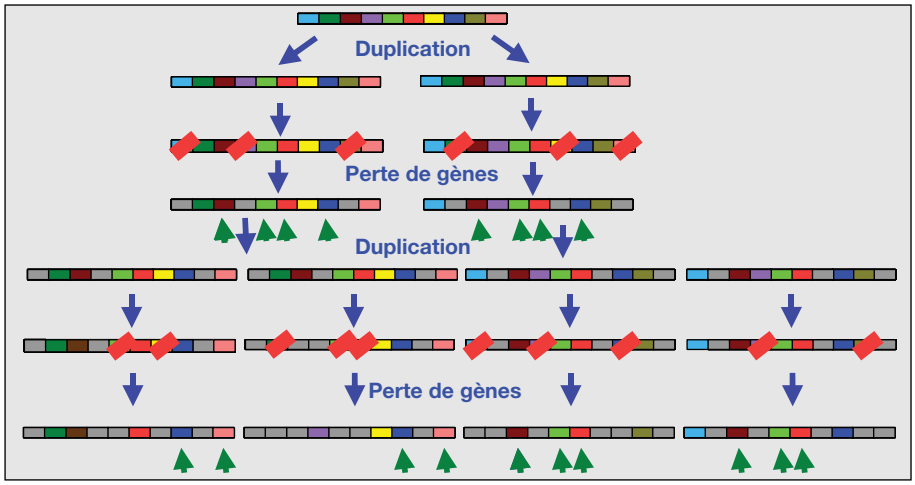

Figure 1. Schéma d'évolution des génomes par duplication et perte de gènes.

Figure 1. Representation of genome evolution through gene duplication and loss.

Chaque gène est représenté par une couleur et devient gris lorsqu'il est inactivé et perdu. Les flèches vertes indiquent les gènes conservés par paires au sein d'une duplication.

impliqués dans les réponses aux stress sont beaucoup moins bien conservés et semblent évoluer plus vite.

D'autres génomes et d'autres duplications

La motivation principale d'études sur des génomes d'espèces modèles est d'en transférer les résultats et de les exploiter pour analyser les génomes d'espèces cultivées. Or, si on savait que les génomes de riz et de graminées étaient largement colinéaires (Moore *et al.*, 1995), on ne savait rien de leur relation avec les autres monocotylédones ni des relations entre *Arabidopsis* et les dicotylédones.

Arabidopsis et les dicotylédones

Avant même que la séquence du génome d'*Arabidopsis* ne soit achevée, différents laboratoires ont essayé de cartographier sur des espèces cultivées de dicotylédones vraies des gènes relativement bien conservés entre *Arabidopsis* et riz : on pouvait s'attendre à ce qu'ils soient aussi relativement bien conservés entre *Arabidopsis* et ces espèces plus proches d'*Arabidopsis* que du riz. Les premiers résultats ont été confus et difficiles à interpréter, jusqu'à ce que l'on réalise l'existence de WGD chez *Arabidopsis* et que l'on se mette à utiliser directement des EST de l'espèce concernée plutôt que celles d'*Arabidopsis*. En utilisant des traitements statistiques rigoureux, il a été possible de décrire des blocs chromosomiques chez chacune des espèces étudiées, dans lesquels les marqueurs étaient dans le même ordre que sur un bloc chromosomique d'*Arabidopsis*. Cette analyse a aussi révélé chez chacune des espèces ciblées l'existence de larges régions dupliquées, probablement elles aussi résultant de WGD (Dominguez *et al.*, 2003). Ainsi, parmi 57 segments du génome d'*Arabidopsis*, 27 sont colinéaires avec des segments des génomes de betterave, 49 avec des segments de pomme de terre, 24 avec ceux du tournesol et 37 avec ceux du pêcher. Seize de ces blocs sont communs à ces cinq espèces et 32 sont partagés par au moins trois d'entre elles. Selon les espèces, ces blocs recouvrent de 16 à 33 % du génome d'*Arabidopsis* (Delseny, 2004). Ces observations initiales ont depuis été confirmées et étendues à d'autres espèces (Blanc et Wolfe, 2004a). On peut donc présumer que ces blocs représentent le génome ancestral des dicotylédones et qu'au moins les WGD les plus anciennes existaient déjà dans le génome ancestral (Bowers *et al.*, 2003). Cependant, il est manifeste que, depuis leur spéciation à partir d'un ancêtre commun, les génomes de dicotylédones ont été fortement remaniés et il est peu probable que l'on puisse faire un usage systématique de la synténie[11] avec *Arabidopsis* pour les analyser. Cette analyse est par ailleurs limitée aux seuls gènes cartographiés chez les espèces cultivées, et elle sous-estime la réalité du fait de l'application de traitements statistiques rigoureux.

Arabidopsis, riz et graminées

La détermination de la séquence du riz, a permis de comparer directement les deux séquences modèles. Cette comparaison a révélé quelques traces de synténie significatives, mais en aucun cas sur des grandes distances (Salse *et al.*, 2002). Elle a aussi permis de comparer précisément le riz avec le maïs, le sorgho et le blé. Ces travaux ont globalement confirmé la colinéarité de ces différents génomes, mais ont aussi révélé un nombre de réarrangements plus élevé qu'anticipé. Ils ont également démontré l'existence de WGD chez le riz, plus anciennes à une exception près, sur les chromosomes 11 et 12, que la WGD la plus récente d'*Arabidopsis*. Vingt-neuf blocs dupliqués représentent environ 72 % du génome du riz (Salse *et al.*, 2004, Yu *et al.*, 2005). Les mêmes blocs dupliqués ont été observés à la fois chez le riz, le maïs, le sorgho et le blé, ce qui indique que cette WGD est antérieure à la différenciation de ces espèces, estimée à environ 50-70 millions d'années. L'ensemble de ces travaux a permis de se faire une idée assez précise de la façon dont les génomes de céréales ont évolué à partir d'un ancêtre commun (Wei *et al.*, 2007 ; Singh *et al.*, 2007 ; Salse *et al.*, 2008 ; Bolot *et al.*, 2008 ; Paterson *et al.*, 2009), sans doute à cinq paires de chromosomes. Cet ancêtre aurait subi une WGD entre 70 et 90 millions d'années pour donner une espèce à 10 paires qui, à la suite de cassures et de fusions chromosomiques, aurait donné des espèces à 12 paires, comme le riz ou son ancêtre direct. Des processus de cassure/fusion similaires, entre 50 et 30 millions d'années, auraient conduit aux ancêtres du blé, à 7 paires de chromosomes, et à ceux à 10 paires du sorgho et du maïs. Sorgho et maïs se sont différenciés vers 12 millions d'années et, finalement, une nouvelle WGD chez le maïs autour de cinq millions d'années, suivie de fusions chromosomiques aurait conduit au génome actuel du maïs. La canne à sucre, *Saccharum officinalis*, constitue

[11] Portions d'organisation identique des gènes sur un même chromosome pour des espèces apparentées.

un exemple extrême de cette évolution, puisque cette espèce présente un génome octoploïde dont l'organisation de base est très proche de celui du sorgho. On ne dispose, pour le moment, d'aucune information sur les monocotylédones autres que les céréales et le séquençage de génomes appartenant à d'autres familles devrait éclairer leur évolution.

Trois nouveaux génomes de dicotylédones disponibles

En 2006, 2007 et 2008, les séquences partielles de trois nouveaux génomes ont été rapportées : le peuplier (Tuskan *et al.*, 2006), la vigne (Jaillon *et al.*, 2007, Velasco *et al.*, 2007) et le papayer (Ming *et al.*, 2008). Leur analyse et leur comparaison ont profondément modifié notre compréhension de l'évolution des génomes végétaux et conduit à une révision de notre concept de plante modèle. La percée vient de l'analyse du génome de la vigne, qui révèle qu'un grand nombre de gènes sont en trois copies réparties sur trois chromosomes ou fragments de chromosomes, suggérant une nature hexaploïde cryptique de ce génome. Par ailleurs, lorsque chacune de ces régions est alignée avec le génome d'*Arabidopsis*, on observe quatre régions correspondantes, ce qui suggère que, depuis la séparation à partir d'un ancêtre commun, deux WGD se sont produites chez *Arabidopsis*, alors qu'il ne semble pas y en avoir eu chez la vigne. Le même type de comparaison avec le peuplier, indique que, depuis la divergence à partir de l'ancêtre commun, le peuplier a subi une WGD, puisqu'à chaque segment chromosomique de la vigne correspondent deux segments de celui du peuplier (Jaillon *et al.*, 2007). La publication du génome du papayer est venue confirmer ce schéma évolutif. Il y a une bonne correspondance entre le génome de la vigne et celui du papayer – chaque fragment génomique du papayer correspond à un fragment unique de celui de la vigne. L'ancêtre commun aux dicotylédones était donc probablement déjà hexaploïde et, à partir de ce dernier, ont divergé des espèces qui ont subi une WGD (peuplier), deux WGD (*Arabidopsis*) ou aucune WGD (vigne et papayer). D'autres espèces, comme les choux ou le colza, ont pu subir des événements de polyploïdisation supplémentaires. L'ensemble de ces observations démontre clairement que les processus de duplication globale des génomes accompagnés de perte ou de spécialisation des gènes dupliqués constituent un moteur puissant et majeur de l'évolution des plantes. Il apparaît ainsi, au bout du compte, que les génomes de la vigne et du papayer ont un génome dont l'organisation est beaucoup plus simple que celle de celui d'*Arabidopsis* et que ce dernier n'est sans doute pas le modèle dont nous pouvions rêver. L'alignement de ces différents génomes permet *a priori* de reconstituer un génome *consensus* pour l'ancêtre des dicotylédones (Tang *et al.*, 2008). En revanche, la comparaison entre le génome de la vigne ou du papayer et le génome du riz ne permet pas de révéler un profil clair de correspondances entre segments chromosomiques, ce qui pourrait suggérer que la structure hexaploïde ancestrale des dicotylédones est postérieure à la séparation des dicotylédones et des monocotylédones.

Le séquençage en cours de nouveaux génomes, en particulier d'espèces à la base des arbres phylétiques, pourrait bien révéler de nouvelles surprises.

Conclusion

Les données accumulées ces 20 dernières années ont déjà considérablement amélioré notre compréhension de l'organisation des génomes et des grandes fonctions biologiques qui contrôlent la vie d'une plante et ses interactions avec l'environnement, en permettant la découverte de gènes clés. Cette phase de découverte n'en est encore qu'à ses tous débuts, car seuls deux génomes ont réellement été analysés en détail, ceux d'*Arabidopsis* et du riz. Il reste encore beaucoup à découvrir sur ces deux espèces, où plusieurs milliers de gènes sont encore sans fonction connue et où l'analyse fonctionnelle est en plein essor. Il en reste encore plus à découvrir pour les autres espèces dont les génomes seront disponibles prochainement.

Les données acquises sur le repérage ou l'identification des gènes d'intérêt ont déjà un impact important sur l'amélioration des plantes avec l'utilisation des méthodologies de sélection assistée par marqueurs. L'autre développement attendu concerne la transgénèse : l'un des facteurs limitant était le peu de gènes caractérisés disponibles pour manipuler les génomes et améliorer les qualités et les performances des espèces cultivées. Cette situation change rapidement. Dans le même temps, les progrès du séquençage permettent de caractériser directement au niveau de la séquence les nouvelles plantes transgéniques (Ming *et al.*, 2008) et augmentent ainsi la sécurité de ces nouvelles variétés. La disponibilité d'un nombre croissant de génomes végétaux amène le développement de la génomique comparée. Dans le même temps, les technologies de séquençage ont évolué et vont révolutionner notre façon d'aborder la biologie et l'amélioration des plantes. Ces perspectives sont développées, dans ce même numéro, par Delseny (2009). ■

Remerciements

L'auteur remercie ces collègues du laboratoire Génomes et développement des plantes (LGDP) et les nombreux autres collègues qu'il n'a pas été possible de citer, faute de place, mais qui ont contribué par leurs travaux à l'essor de la génomique végétale.

Références

Adams MD, Kelley JM, Gocayne JD, *et al.* Complementary DNA sequencing : expressed sequence tags and human genome project. *Science* 1992 ; 252 : 1651-6.

Alonzo FJM, Stepanova AN, Leisse TJ, *et al.* Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* 2003 ; 301 : 653-7.

Ayele M, Haas BJ, Kumar N, *et al.* Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in Arabidopsis. *Genome Res* 2005 ; 15 : 487-95.

Baerenfaller K, Grossmann J, Grobei MA, *et al.* Genome-scale proteomics reveals Arabidopsis gene models and proteome dynamics. *Science* 2008 ; 320 : 938-41.

Barakat A, Miranda-Szick K, Chang IF, *et al.* The organization of cytoplasmic ribosomal protein genes in the Arabidopsis genome. *Plant Physiol* 2001 ; 127 : 398-415.

Ben Amor B, Wirth S, Merchan F, *et al.* Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. *Genome Res* 2009 ; 19 : 57-69.

Berardini TZ, Mundodi S, Reiser L, *et al.* Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol* 2004 ; 135 : 745-55.

Bevan M, Bancroft I, Bent E, *et al.* The EU Arabidopsis Genome Project : Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. *Nature* 1998 ; 391 : 485-8.

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* 2000 ; 12 : 1093-101.

Blanc G, Hokamp K, Wolfe KH. A recent polyploidy superimposed on older large scale duplication in the Arabidopsis genome. *Genome Res* 2003 ; 13 : 137-44.

Blanc G, Wolfe KH. Widespread paleoploïdy in model plant species inferred from age distributions in duplicate genes. *Plant Cell* 2004 ; 16 : 1667-78.

Blanc G, Wolfe KH. Functional divergence of duplicated genes formedby polyploidy during Arabidopsis evolution. *Plant Cell* 2004 ; 16 : 1679-91.

Bolot S, Abrouk M. Masood-Quraishi, Stein N, Messing J, Feuillet C, Salse J. The "inner circle" of the cereal genomes. *Cur Opin Plant Biol* 2008 ; 12 : 1-7.

Bowers JE, Chapman BA, Rong J, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosome duplication events. *Nature* 2003 ; 422 : 433-8.

Cooke R, Raynal M, Laudié M, *et al.* Further progress towards a catalogue of all Arabidopsis genes: analysis of a set of 5000 non redundant ESTs. *Plant J* 1996 ; 11 : 1127-40.

Cooke R, Piegu B, Panaud O, *et al.* From rice to other cereals: comparative genomics. In : Uppadyahia N, Dennis E, eds. *Rice functional genomics*. Heidelberg : Springer-Verlag, 2007.

Delseny M. Re-evaluating the relevance of ancestral shared synteny as a tool for crop improvement. *Curr Opin Plant Biol* 2004 ; 7 : 126-31.

Delseny M, 2009. Le séquençage des génomes de plantes : vers une nouvelle révolution en biologie végétale. *Cah Agri* 2009 ; 6 : epub. Doi: 10.1684/agr.2009.0342

Dominguez I, Graziano E, Gebhardt C, *et al.* Plant genome archaeology: evidence for conserved ancestral chromosome segments in dicotyledonous plant species. *Plant Biotech J* 2003 ; 1 : 91-9.

Goff SA, Ricke D, Lan TH, *et al.* A draft sequence of the rice genome (*Oryza sativa* L ssp *japonica*). *Science* 2002 ; 296 : 92-100.

Gregory BD, Yazaki J, Ecker JR. Utilizing tiling microarrays for whole –genome analysis in plants. *Plant J* 2008 ; 53 : 636-44.

Hofte H, Desprez T, Anselem J, *et al.* An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNA from *Arabidopsis thaliana*. *Plant J* 1993 ; 4 : 1051-61.

Jaillon O, Aury JM, Noel B, *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007 ; 449 : 463-7.

Ilic L, Kellogg EA, Jaiswal P, *et al.* The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. *Plant Physiol* 2007 ; 143 : 587-99.

International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* 2005 ; 436 : 793-800.

Kasschau KD, Fahlgren N, Chapman EJ *et al.* Genome –wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biology* 2007; 5 : e57. Doi:10.1371/journal.p.bio.0050057.

Lee JY, Levesque M, Benfey PN. High-throughput RNA isolation technologies. New tools for high resolution gene expression profiling in plant systems. *Plant Physiol* 2005 ; 138 : 585-90.

Ming R, Hou S, Feng Y, *et al.* The draft genome of the transgenic tropical fruit tree Papaya (*Carica papaya* Linnaeus). *Nature* 2008 ; 452 : 991-6.

Moore G, Devos K, Wang Z, Gale M. Grasses, line up and form a circle. *Curr Biol* 1995 ; 5 : 737-9.

Newman T, De Bruijn FJ, Green P, *et al.* Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. *Plant Physiol* 1994 ; 106 : 1241-55.

Paterson AH, Bowers JE, Bruggmann R, *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 2009 ; 457 : 551-6.

Sallaud C, Gay C, Larmande P, *et al.* High throughput T-DNA insertion mutagenesis in rice : a first step towards in silico reverse genetics. *Plant J* 2004 ; 39 : 450-64.

Salse J, Piegu B, Cooke R, Delseny M. 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* 2002 ; 30 : 2316-28.

Salse J, Piegu B, Cooke R, Delseny M. New *in silico* insight into the synteny between rice (*Oryza sativa* L) and maize (*Zea mays* L.) highlights reshuffling and identifies new duplications in the rice genome. *Plant J* 2004 ; 38 : 396-409.

Salse J, Bolot S, Throude M, *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* 2008 ; 20 : 11-24.

Sasaki T, Song J, Koga-Ban Y, *et al.* Toward cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. *Plant J* 1994 ; 6 : 615-24.

Seki M, Naruska M, Kamiya A, *et al.* Functional annotation of full length Arabidopsis cDNA collection. *Science* 2002 ; 296 : 141-7.

Singh NK, Dalal V, Batra K, *et al.* Single copy genes define a conserved order between rice and wheat for understanding differences caused by duplication, deletion and transposition of genes. *Funct Integr Genomics* 2007 ; 7 : 17-35.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson A. Synteny and collinearity in plant genomes. *Science* 2008 ; 320 : 486-8.

The Arabidopsis Genome Initiative. Sequence and analysis of the flowering plant *Arabidopsis thaliana*. *Nature* 2000 ; 408 : 796-815.

Tuskan GA, DiFazio S, Jansson S, *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr.& Gray). *Science* 2006 ; 313 : 1596-604.

Velasco R, Zharkihk A, Trojio M, *et al.* A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PloS One* 2007 ; 2 : e13-26.

Wei F, Coe E, Nelson W, *et al.* Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genetics* 2007 ; 3 : 1254-63.

Yu J, Hu S, Wang J, *et al.* A draft sequence of the rice genome (*Oryza sativa* L ssp *Indica*). *Science* 2002 ; 296 : 79-92.

Yu J, Wang J, Lin W, *et al.* The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol* 2005 ; 3 : 266-81.

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W. GENEVESTIGATOR: Arabidopsis microarray database and analalysis toolbox. *Plant Physiol* 2004 ; 136 : 2621-32.

Zhou X, Sunkar R, Jin H, *et al.* Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in *Oryza sativa*. *Genome Res* 2009 ; 19 : 70-8.