m/s MINI-SYNTHÈSE
médecine/sciences 2000 ; 16 : ???-?

# Le génome de la mouche du vinaigre

Il y a 90 ans, Thomas H. Morgan isolait la mutation *white* chez la drosophile, amorçant des recherches qui allaient le conduire à formuler les bases de la théorie chromosomique de l'hérédité. Aujourd'hui encore, la puissance de la génétique a fait de ce diptère un modèle de choix pour des recherches dans des domaines aussi variés que la biologie cellulaire, le développement, le comportement ou la neurogenèse. Très récemment, le génome de la drosophile a été séquencé par la société Celera genomics en collaboration avec des institutions académiques américaines et européennes (le *« Berkeley Drosophila Genome Project »*, et l' *« European Drosophila Genome Project »*). Les résultats de ce projet d'envergure qui ont été publiés dans le journal *Science* [1-3] auront un large impact sur la communauté scientifique pour deux raisons. D'une part, les méthodes utilisées tant pour le séquençage que pour l'annotation du génome de la drosophile marqueront un tournant dans les approches de génomique. D'autre part, après la levure et le nématode, la drosophile est le troisième eucaryote dont le génome est entièrement séquencé.

## Le séquençage aléatoire global du génome de la drosophile

Pour la première fois, la technique de séquençage aléatoire *(shot gun)* global a été appliquée à un organisme possédant un génome de grande taille. Cette technique consiste à séquencer de manière aléatoire l'ensemble du génome, sans réaliser préalablement le fastidieux travail de cartographie. Les séquences obtenues sont ensuite assemblées et ordonnées à l'aide d'un programme informatique [2, 4]. Cette approche avait été appliquée avec succès par Graig Venter sur le génome d'une bactérie de petite taille [5]. En mai 1998, Craig Venter et le groupe Perkin Elmer, spécialisé dans la fabrication de séquenceurs, annonçaient leur intention de séquencer l'ensemble du génome humain par cette approche. Une entreprise privée, *Celera genomics*, concentrant d'imposants moyens de séquençage et d'informatique, a été créée dans ce but. Le séquençage de la drosophile a servi de test préliminaire à cette stratégie avant de l'appliquer sur le génome humain. Il aura fallu moins de six mois à cette compagnie pour séquencer, avec une couverture de 10X, le génome de la drosophile. Les séquences obtenues ont été ensuite ordonnées à l'aide de programmes informatiques [2]. En intégrant des données déjà connues ou obtenues indépendamment par d'autres centres de séquençage, un assemblage contenant l'essentiel des régions codantes (partie du génome appelée l'euchromatine) de la drosophile, soit environ 120 mégabases, a été réalisé. Soulignons que la séquence du génome de la drosophile n'est pas complète. Il reste plus d'un millier de « trous » (généralement de quelques kilobases) à combler. Finalement, l'hétéchromatine, qui est composée de séquences répétées non codantes et qui comprend 60 mégabases des 180 mégabases du génome de la drosophile, est inaccessible, quelles que soient les stratégies de séquençage utilisées. En comparant les données obtenues dans une région extrêmement bien connue (2,9 mégabases autour du gène codant pour l'alcool déshydrogénase), les auteurs concluent que l'approche de séquençage aléatoire global développée par Celera donnerait des résultats semblables aux approches traditionnelles utilisant la cartographie, validant de fait cette stratégie pour l'analyse de génomes de grande taille [2]. Considérant qu'il est impossible de prétendre à un séquençage complet d'un génome eucaryote complexe, les auteurs proposent une liste de paramètres permettant d'établir les critères de finition d'un génome : nombre de gènes connus retrouvés, taux d'erreurs de séquence, nombre de « trous »… [1].

Une fois la séquence complète établie, commence la seconde phase de l'analyse, appelée « annotation », qui consiste à déduire de la séquence l'organisation des gènes. A la différence des génomes eucaryotes précédemment analysés (levure, nématode), l'organisation des gènes présents dans le génome de la drosophile est complexe à cause de l'existence de longues régions intergéniques, de la présence de nombreux introns, et du recours fréquent à l'épissage alternatif. Ainsi, les 120 mégabases d'euchromatine contiendraient 24,1 mégabases de séquence codante et 20 mégabases d'introns (allant de 40 paires de base à 70 kilobases). La densité de gènes, qui fluctue de 1 à 30 gènes pour 50 kilobases, est en moyenne d'un gène tous les 9 kilobases chez la drosophile. Cette organisation complique singulièrement le repérage des gènes présents dans le génome de la drosophile. L'annotation consiste dans un premier temps à prédire, à l'aide de programmes informatiques, l'existence des gènes ; puis dans un second temps, un « annotateur » se fondant sur des données extérieures (présence de transcrits connus, séquence codant pour des peptides ayant des homologies avec des protéines présentes dans les bases de données…) décide, selon des critères heuristiques, de l'existence ou de l'absence de ce gène. Une partie importante de l'annotation du génome de la drosophile a été réalisée au cours d'un « jamboree », regroupant des spécialistes de l'annotation et des généti-

ciens de la drosophile travaillant sur des thématiques variées et apportant chacun une expertise dans un domaine. Cette réunion informelle et intense a permis de réaliser une première annotation avec une remarquable efficacité et de tirer du génome un certain nombre d'informations. Cela étant, on peut estimer que de nombreux gènes ne sont pas correctement annotés. Ce travail de bénédictin sera poursuivi avec le reste de la communauté des généticiens de la drosophile. Après tout, si l'on estime à 5 000 le nombre de chercheurs travaillant sur la drosophile, cela ne fera que deux gènes et demi par personne ! Ce projet de longue haleine sera facilité par l'existence d'une base de données faisant autorité et regroupant l'ensemble des données sur le génome de la drosophile (FLYBASE http://flybase.bio.indianaedu:82/).

Le séquençage du génome de la drosophile aura donc été réalisé et publié en seulement 10 mois ! Ce « coup de pouce » lié à l'implication de la société Celera dans le séquençage ne peut que réjouir les spécialistes de la drosophile, qui n'en demandaient pas tant d'une compagnie privée. Ainsi, dans la course au séquençage du génome humain que se livrent les institutions publiques et la compagnie Celera, les « drosophilistes » ont pu tirer leur épingle du jeu. Il est probable que les stratégies de séquençage mises en jeu pour la drosophile auront un impact important sur l'analyse du génome d'autres organismes (homme, souris…).

**Les révélations du génome de la drosophile**

Une première surprise est le faible nombre de gènes, 13 601, codés par le génome de la drosophile *(Tableau I)*. Il faut toutefois noter que ce nombre reste très approximatif puisque chez le nématode *C. elegans*, pas moins de 1 000 nouveaux gènes ont été identifiés depuis la publication de la séquence. Ainsi, *C. elegans*, possède un plus grand nombre de gènes que la drosophile (18 424), bien qu'il présente un nombre limité de cellules et une moins grande complexité apparente. Cependant, si l'on regarde de plus près, ces différences s'expliquent essentiellement par l'extension de certaines familles de gènes ; ainsi, le nématode contient mille copies d'un gène codant pour un récepteur olfactif qui n'est présent qu'en soixante exemplaires chez la drosophile. Si l'on compare la diversité des protéines, on note que les génomes de la drosophile et du nématode codent pour un nombre similaire de familles différentes de protéines *(Tableau 1)*, lequel n'est que deux fois plus élevé que le nombre observé chez la levure, un champignon unicellulaire. Finalement, plus de 50 % des protéines de la drosophile présentent de fortes similarités avec des protéines de l'homme contre 36 % chez le ver [3]. Une comparaison du génome de drosophile, du nématode et de la levure indique que le nématode et la drosophile possèdent un nombre significativement plus grand de protéines contenant des domaines multiples, comparé à la levure, la plupart de ces protéines étant des récepteurs transmembranaires avec des domaines extracellulaires impliqués dans la reconnaissance.

L'utilisation d'un système modèle accessible à l'analyse génétique est devenue un recours classique pour mieux comprendre les bases moléculaires de certaines maladies génétiques de l'homme [6]. Sur 289 gènes dont l'altération est la cause de graves pathologies chez l'homme, 177 ont un véritable homologue chez la drosophile (soit 61 %, ce nombre étant vraisemblablement sous-évalué). L'absence de certains gènes s'explique aisément par des différences physiologiques entre la drosophile et les mammifères. Ainsi, on ne trouvera aucun homologue de l'hémoglobine chez la drosophile dont le transport de l'oxygène est assuré par un système de trachées, même constat pour les gènes codant pour les enzymes impliquées dans le réarrangement des immunoglobulines. De même, 68 % des gènes impliqués dans des cancers chez l'homme ont une contrepartie chez la drosophile. Ainsi, la drosophile possède bien un homologue de p53 mais pas des gènes *BRCA1* et *BRCA2*. A quelques exceptions près, lorsqu'un homologue humain est trouvé chez la drosophile, il est aussi présent dans le génome de *C. elegans*. Le séquençage des génomes de la drosophile, du nématode et de la levure facilitera considérablement la recherche d'un système modèle accessible à la génétique. Il ne restera plus, pour les chercheurs, qu'à sélectionner lequel parmi ces trois organismes est le plus pertinent au regard de la pathologie étudiée.

L'étude des voies de transduction du signal est un domaine particulièrement dynamique chez la drosophile. Un grand nombre de ces cascades sont conservées entre la drosophile et les mammifères (TGFβ, récepteur

Tableau I

COMPARAISON DES GÉNOMES D'*HAEMOPHILUS INFLUENZAE* (BACTÉRIE), DE *SACCHAROMYCES CEREVISIAE* (LEVURE), DE *CAENORHABDITIS ELEGANS* (NÉMATODE) ET DE *DROSOPHILA MELANOGASTER* (MOUCHE)

| | ***H. influenzae*** **(1,83 Mb)** | ***S. cerevisiae*** **(12,1 Mb)** | ***C. elegans*** **(97 Mb)** | ***D. melanogaster*** **(180 Mb)** |
|---|---|---|---|---|
| Nombre total de gènes prédits | 1 709 | 6 241 | 18 424 | 13 601 |
| Nombre de gènes dupliqués | 284 | 1 858 | 8 971 | 5 536 |
| Nombre de familles de protéines | 1 425 | 4 383 | 9 453 | 8 065 |

*La ligne 1 indique le nombre total de gènes pour chacune des espèces considérées. La ligne 2 indique le nombre de gènes de chacune des espèces qui seraient issus d'une duplication. La ligne 3 indique le nombre de familles de gènes codant pour des protéines différentes. La taille du génome est indiquée en mégabase sous le nom de l'organisme (d'après [3]).*

à tyrosine kinase, Notch/lin12, Toll/NF-κB, Jak/STAT, Hedgehog). L'étude du génome de la drosophile révèle, qu'à l'instar des mammifères, les protéines impliquées dans la signalisation appartiennent souvent à de larges familles : 8 gènes codent pour des récepteurs Toll, 7 pour des protéines de la famille wingless ou TGFβ. En revanche, à la différence des vertébrés, les gènes codant pour Notch, Hedgehog ou Jak sont présents en une seule copie.

L'analyse du génome de la drosophile a permis d'autres révélations inattendues. Si le ver et la drosophile possèdent respectivement 260 et 450 peptidases, cette différence s'explique uniquement par l'expansion d'un type de peptidase, les protéases à sérine au nombre de 199 chez la mouche du vinaigre contre 7 chez le ver [3]. Or, moins d'une dizaine de ces protéases, jouant un rôle dans l'établissement de l'axe dorsoventral de l'embryon, ont fait l'objet d'études approfondies chez la drosophile. En fait, rien n'est connu des cascades de protéases mises en jeu dans la coagulation, l'immunité, ou la signalisation et le grand nombre de gènes codant pour des protéases à sérine suggère que la drosophile est un bon modèle d'étude des cascades protéolytiques.

De même, de très nombreux gènes codant pour des protéines non encore identifiées chez la drosophile mais possédant, dans d'autres organismes, des homologues impliqués dans la neurogenèse, les régulations hormonales, ou le métabolisme ont été identifiés au cours de ce projet. Il est probable que les données issues du séquençage vont jouer un rôle de catalyseur pour des projets de recherches portant sur la physiologie et le métabolisme de la drosophile, domaines qui ont été délaissés par les généticiens au profit d'autres thématiques comme l'étude des gènes du développement.

**Conclusions**

Les données obtenues lors du séquençage du génome de la drosophile constituent un atout de plus pour ce système modèle. Elles vont considérablement faciliter l'identification des gènes isolés lors d'un criblage génétique. Elles vont aussi permettre de dynamiser les recherches dans des thématiques émergentes ou négligées. Le dialogue stimulant qui existe entre spécialistes des vertébrés et généticiens de la drosophile devrait s'en trouver renforcé. Libre à chacun de regarder si son gène favori est présent chez la drosophile et, pourquoi pas, si par chance, il n'existe pas déjà une lignée mutante pour ce gène… Au dernier congrès sur la drosophile qui s'est tenu en mars dernier à Pittsburgh (États-Unis), des travaux ont fait état de récents progrès qui devraient permettre bientôt d'inactiver un gène par recombinaison homologue, seule technique qui manque encore à l'arsenal des généticiens de la drosophile. On peut aussi espérer obtenir, dès la fin de l'année 2000, les premières « puces à ADN » [6] contenant tous les gènes de la drosophile, ouvrant ainsi la possibilité d'étudier le profil transcriptionnel global.

Finalement, il est utile de rappeler que la drosophile appartient à l'ordre des insectes, lequel rassemble plus de 90 % des espèces animales. A l'heure où les techniques de transgenèse connaissent leurs premiers succès chez d'autres insectes (ver à soie, moustique…), la connaissance du génome de la drosophile permettra une meilleure compréhension d'un groupe prolifique qui partage avec les mammifères une réussite évolutive certaine.

## RÉFÉRENCES


1. Adams MD, Celniker SE, Holt RA, *et al.* The genome sequence of drosophila melanogaster. *Science* 2000 ; 287 : 2185-95.

2. Myers EW, Sutton GG, Delcher AL, *et al.* A whole-genome assembly of Drosophila. *Science* 2000 ; 287 : 2196-204.

3. Rubin GM, Yandell MD, Wortman JR, *et al.* Comparative genomics of the eukaryotes. *Science* 2000 ; 287 : 2204-15.

4. Weissenbach J, Salanoubat M. Séquence des génomes : le feu d'artifice. *Med Sci* 2000 ; 16 : 10-6.

5. Fleishmann RD, Adam MD, White O, *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995 ; 269 : 496-512.

6. Birman S. La drosophile, un modèle génétique pour l'étude des maladies neurodégénératives. *Med Sci* 2000 ; 16 : 164-70.

7. Lee P, Hudson HJ. La puce à ADN en médecine et en science. *Med Sci* 2000 ; 16 : 43-9.



**Bruno Lemaître**

*Centre de génétique moléculaire, Bâtiment 26, Cnrs, 91198 Gif-sur-Yvette, France.*