Ce que nous apprend l'analyse des génomes au sujet de l'évolution

Jacques van Helden

Chargé de cours à l'Université Libre de Bruxelles
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/Users/jvanheld/
Jacques.van.Helden@ulb.ac.be

Résumé

Depuis 1995, nous disposons des séquences génomiques complètes pour un nombre sans cesse croissant d'espèces microbiennes, animales et végétales. La première leçon de l'analyse des génomes a été la mesure de notre ignorance : plus d'un tiers des gènes identifiés dans le génome de la levure étaient totalement inconnus, et on n'avait aucune idée de leur fonction. La connaissance des génomes complets d'un grand nombre d'organismes ouvre des perspectives nouvelles pour la compréhension des mécanismes de l'évolution. Certains y voient la perspective de passer du « siècle du gène » (selon l'expression d'Evelyn Fox-Keller) au « siècle du système ». La modélisation d'un génome entier reste cependant, de très loin, hors de notre portée. Par contre, la génomique comparative permet déjà de poser des questions qui auraient été inconcevables en se limitant à une étude de gènes individuels. On peut par exemple retracer les événements évolutifs tels que réarrangements chromosomiques, duplications de régions génomiques localisées ou de génomes entiers, transferts horizontaux. Les génomes complets permettent d'analyser non seulement la présence mais également l'absence de gènes à travers la taxonomie, et détecter ainsi des modules de gènes fonctionnellement liés. On peut également analyser l'obsolescence des gènes, leur « érosion » progressive dans certaines espèces où leur fonction n'est plus requise.

L'ère de la génomique

La génomique (analyse des génomes) est une extension de la génétique moléculaire, dont l'objet d'étude passe du gène (en tant que support physique d'un caractère observable) au génome entier (ensemble de l'information génétique d'un organisme). Si le 20ème siècle a pu être qualifié de « siècle du gène » (Fox-Keller, 2003), le 21ème siècle s'ouvre pour les biologistes à l'ère de la génomique. Nous disposons aujourd'hui de la séquence génomique complète de plusieurs centaines d'organismes unicellulaires (bactéries, levures, protozoaires), de quelques dizaines d'animaux pluricellulaires (humain, souris, rat, chien, ornithorynque, poulet, mouche, etc.), et de quelques plantes.

La question que nous abordons ici est de savoir en quoi ce passage à l'échelle des génomes peut contribuer à notre compréhension des mécanismes de l'évolution.

La théorie darwinienne de l'évolution se base sur l'interaction entre deux facteurs essentiels : les « variations » qui apparaissent de façon « accidentelle » (au sens d'aléatoire) entre individus d'une même espèce, et la sélection naturelle. Depuis sa publication, la théorie de Darwin a été précisée et renforcée par un siècle et demi de découvertes en biologie. Un tournant important a été l'intégration de diverses disciplines liées à la biologie (paléontologie, systématique, génétique, génétique des populations), pour donner naissance à ce que nous appelons la théorie synthétique de l'évolution. Plus récemment, la génétique moléculaire a sans conteste apporté une contribution significative à la compréhension des mécanismes évolutifs : l'analyse des séquences permet d'établir le lien entre les mutations et le phénotype (la manifestation de ces mutations sous la forme de variations interindividuelles).

Jusqu'au début des années 1990, le séquençage de l'ADN représentait un travail important pour les biologistes moléculaires. Un doctorant pouvait passer une partie significative de sa thèse à séquencer quelques kilobases afin de caractériser un seul gène. Les projets de séquençage de génomes qui ont vu le jour au début des années 1990 ont stimulé le développement de méthodes de séquençage automatique, et mené à des progrès technologiques extrêmement rapides. Le nombre de séquences disponibles dans les bases de données augmente de façon exponentielle depuis une dizaine d'années, et ce mouvement va encore s'amplifier avec la nouvelle génération de machines à séquencer.

La composition des génomes

La Table 1 fournit quelques ordres de grandeur concernant la taille et le contenu global des génomes. Les premiers génomes publiés étaient ceux d'organismes microbiens (bactéries, levures), car ils étaient de taille raisonnable pour les méthodes de séquençage disponibles à l'époque. Ces génomes sont très compacts: un génome bactérien typique compte quelques millions de paires de bases, dont la majorité (environ 85%) correspond à des régions codantes. Ces génomes contiennent en moyenne un gène codant par kilobase[1]. Chacun de ces gènes contient l'information (le code) permettant de synthétiser une protéine (ou dans certains cas une sous-unité d'un complexe protéique). Les protéines sont les « acteurs moléculaires » qui assurent les fonctions essentielles au fonctionnement de la cellule : les enzymes cataboliques permettent de dégrader les substances nutritives du milieu ; les enzymes anaboliques sont impliquées dans la synthèse des petites molécules intervenant à tous les niveaux cellulaires ; les transporteurs assurent les échanges de nutriments entre la cellule et son milieu; les protéines de régulation permettent à l'organisme de répondre à des

[1] Dans le cadre de ce chapitre, nous parlerons essentiellement des gènes « codants » (ceux dont la séquence sert de modèle pour la synthèse de protéines), qui représentent la grande majorité des gènes d'un génome. N'oublions cependant pas qu'il existe également des gènes non codants, qui produisent les ARN de transfert et ARN ribosomaux. La définition du gène sera discutée dans l'encadré « Le piège des mots » de cet ouvrage.

modifications de son environnement en activant ou en réprimant l'expression des autres gènes.

Nom d'espèce	Nom commun	Année de publication	Taille du génome Mb	Nombre de gènes	Distance moyenne entre gènes Kb	Fraction couverte par des gènes codants %	Fraction non-codante %	Fraction répétitive %	Fraction transcrite %	Remarques
Bactérie										
Mycoplasma genitalium	*Mycoplasma*	1995	0.6	481	1.2	90	10			Petit génome (intracellulaire)
Haemophilus influenzae		1995	1.8	1 717	1.0	86	14			Premier génome bactérien séquencé
Escherichia coli	*Entérobactérie*	1997	4.6	4 289	1.1	87	13			
Levures										
Saccharomyces cerevisiae	*Levure du boulanger*	1996	12	6 286	1.9	72	28			Premier génome eucaryote
Animaux										
Caenorhabditis elegans	*Ver nématode*	1998	97	19 000	5	27	73			Premier génome de métazoaire
Drosophila melanogaster	*Mouche à vinaigre*	2000	165	16 000	10	15	85			
Ciona intestinalia			174	14 180	12					
Danio rerio	*Poisson zèbre*		1 527	18 957	81					
Xenopus laevis	*Xénope (amphibien)*		1 511	18 023	84					
Gallus gallus	*Poule*		2 961	16 736	177					
Ortnithorynchus anatinus	*Ornithorynque*		1 918	17 951	107					
Mus musculus	*Souris*	2002	3 421	23 493	146					
Pan troglodytes	*Chimpanzé*		2 929	20 829	141					
Homo sapiens	*Humain*	2001	3 200	21 528	149	2	98	46	28	Version "brouillon"
1000 génomes humains		> 2008								Projet annoncé en janvier 2008
Plantes										
Arabidiopsis thaliana	*Arabette*	2001	120	27 000	4	30	70			Premier génome de plante
Oryza sativa	*Riz*		390	37 544	10					
Zea mais	*Maïs*		2 500	50 000	50			50		Nb de gènes approximatif
Triticum aestivum	*Blé*		16 000							Génome hexaploïde
Lilium	*Lys*		120 000							
Psilotum nudum			250 000							

Table 1. Taille des génomes de quelques organismes modèles.

Chez les organismes pluricellulaires, on constate une augmentation drastique de la taille des génomes. Pourtant, le nombre de gènes n'augmente pas de façon proportionnée. L'ADN du génome humain compte 3 milliards de paires de bases, soit 1000 fois plus qu'un génome bactérien moyen, mais ne contient que 10 fois plus de gènes (~25.000). La densité de gènes est donc 100 fois plus faible chez les mammifères que chez les bactéries. La vaste majorité de notre génome est constituée de régions non-codantes (98%). Une partie de ces séquences non codantes contiennent des signaux de régulation, reconnus par les facteurs transcriptionnels qui contrôlent l'expression des gènes. Chez les pluricellulaires, la régulation de l'expression est beaucoup plus complexe que chez les unicellulaires : les gènes sont exprimés selon un pattern spatio-temporel extrêmement précis, qui détermine la différenciation de chaque cellule lors du développement embryonnaire et le fonctionnement des cellules chez l'individu adulte. Nos cellules hépatiques n'expriment pas les mêmes gènes que nos cellules rétiniennes ou cérébrales. Très récemment, on a également réalisé qu'une partie du génome était transcrite pour former des petits ARN non codants. Ces micro-ARN modulent l'activité des ARN messagers, et interviennent donc dans la régulation transcriptionnelle et post-transcriptionnelle de la synthèse des protéines (voir l'article de Jean Vandenhaute « *Gène , épigène et évolution* » dans cet ouvrage).

Un génome ne se limite donc pas, loin s'en faut, à une collection de gènes codant des protéines. Même le plus simple des génomes bactériens correspond à un système biologique dont la complexité dépasse complètement notre niveau actuel de connaissance et d'analyse.

Décryptage des génomes : la mesure de notre ignorance

La Figure 1 montre la séquence d'un fragment de 1000 nucléotides du génome humain. Le génome complet, 3 millions de fois plus long, peut être facilement téléchargée à partir de plusieurs bases de données publiques.

```
....CGATGCTCAAACATTTCAATTTTTTAGGTCAAAAATGCCTTAGGTTTAGCACAGCAATGT
AGGTGCCAAACTCATCGCAGTGAATTGCAGGCGGGAGCAACAAGGACGCCTGCCTCCTTTCTGC
CTGCTTTTTGCAATAGTCCGATTTGAGAAGGGGACCCACGAGAGACACAAAATGCACGCCCCCA
CGCCACATCCTTTTTACCCCGCAATGGGTTAAGACTGTCAACAGGCAGGCCACCTCGCAGCGTC
CGCGGAGTTGCAGGCCCGCCCCCGCCAGGGTGTGGCGCTGTCCCCCTGGCGCTGGGCGGGGGAG
GAGGGGCGCGCGGCGGCCGAGGAGGGGCGCGCGGCGGCCGGGCGGGGCGAGCGGAGGCGAGTGG
AGGACGCGTAGACGCGCCGCGGTCCCCGCCTGCCGCTGCTCCGCCGCAGTCGCCGCTCCAGTCT
ATCCGGCACTAGGAACAGCCCCGAGCGGCGAGACGGTCCCCGCCATGTCTGCGGCCATGAGGGA
GAGGTTCGACCGGTTCCTGCACGAGAAGAACTGCATGACTGACCTTCTGGCCAAGCTCGAGGCC
AAAACCGGCGTGAACAGGAGCTTCATCGCTCTTGGTGGGTGGCCGGGGGTCGCCGCCGCTGGTA
GGGCCACGGGAGCCGCCGCTGCCCCAGCTGCTGGGGAAGGAAGCAGGGAGAGGACTCGGGAAAG
GTGGAGTCGGAGACAGACGGGACAAGCAGCATATTCAGGGATCAGGCTGGCCTCCCGGAAAGCG
TGGGCATCGGAGGACCCCGCGGGGGCTGCCCAGGCTGAGGGTCGCGGGGCTGGAGGGCAGCTGC
GGCGCCGGGCGCTGGCAGCTGGAAGGGCCAGCGCTGACGTATGTCTGCCCCGCGGCCCGGCGCC
CTATTCCTGCTGTCCTGCGCGGTGGGCGCGGACGGCGGGGCCCCTGCGGGCGGGCGCGTTGACG
GAGGTACCCGGTCCTACCCGACCCTCCGTGGAGCTCCGCCCGGAG....
```

Figure 1: fragment de 1000 nucléotides du génome humain.

Le premier problème quand on dispose de « séquences au kilomètre » est de localiser les régions codantes[2], ce qui représente déjà un défi pour les biologistes et les bioinformaticiens. Dans les génomes bactéries, la localisation des régions codantes est relativement aisée : le principe de base consiste à détecter de longues régions dépourvues de codon stop (phases ouvertes de lecture, aussi appelées *ORF* pour « *open reading frames* »), qui correspondent généralement à des régions codantes. Dès qu'il s'agit d'organismes eucaryotes l'organisation des génomes devient beaucoup plus complexe : l'ARN résultant de la transcription (ARN primaire) subit un processus appelé épissage, au cours duquel certains fragments (les introns) sont éliminés, tandis que les autres (les exons) sont rattachés bout à bout pour former l'ARN messager « mature » qui sortira du noyau et permettra la synthèse d'une protéine. La région codante d'un gène ne correspond dont plus à une phase ouverte de lecture : elle peut se trouver dispersée sur plusieurs fragments génomiques séparés par des introns dont la taille peut atteindre des dizaines de kilobases. Sur la seule base de la séquence génomique d'un organisme eucaryote (et particulièrement pour les pluricellulaires), il est extrêmement difficile de détecter les régions codantes.

La difficulté ne s'arrête pas là : après avoir plus ou moins réussi à prédire les régions codantes, il reste à déterminer la fonction des gènes correspondants. Au moment de la

[2] Ainsi que les régions correspondant à des gènes non codants (ARN de transfert, ARN ribosomal, …), que nous ignorons ici pour simplifier l'exposé.

publication du génome de la levure du boulanger, plus de 40% des gènes qu'on y détectait étaient complètement inconnus. Pourtant, cette levure avait servi pendant un demi-siècle d'organisme modèle aux généticiens et aux biologistes moléculaires. Ne parlons même pas d'un génome comme celui de *Plasmodium falciparum*, le vecteur de la malaria, dont on ignore encore la fonction de plus de 60% des gènes, 7 ans après la publication de sa séquence complète.

Une des premières leçons de la génomique est donc de nous inciter à la modestie : l'analyse des premiers génomes disponibles nous a avant tout donné la mesure de notre ignorance. En dépit des annonces quelque peu triomphalistes qui ont accompagné la publication du génome humain, nous sommes très loin d'avoir décrypté le moindre génome, et les promesses de pouvoir guérir les maladies génétiques restent des vœux pieux. Le décryptage des génomes représente un défi formidable, qui mobilisera encore les efforts enthousiastes de centaines de laboratoires aux quatre coins de la planète pendant plusieurs décennies.

Il serait cependant exagérément pessimiste de s'arrêter ici. Ayant pris la pleine mesure des limitations de notre compréhension des génomes, il n'en reste pas moins que les séquences génomiques nous apportent déjà un éclairage nouveau sur les processus et sur les événements évolutifs. Dans le sections suivantes, nous décrirons quelques-unes de ces avancées.

La génomique comparative révèle les réarrangements chromosomiques

Même en ignorant largement la localisation et la fonction des gènes, le fait de disposer de nombreux génomes peut déjà nous apporter des informations précieuses concernant les événements évolutifs du passé. La génomique a stimulé le développement de nouvelles approches informatiques, qui permettent de comparer (d' « aligner ») les génomes complets de deux organismes, et de détecter les similarités et différences. À titre d'exemple, la Figure 2 montre la correspondance entre le chromosome 1 de l'humain, et les chromosomes du chimpanzé (image de gauche) et du poulet (image de droite). Les rectangles colorés indiquent des blocs dont l'organisation est conservée (blocs de synténie). Les flèches indiquent les correspondances entre régions de synténie de l'humain et de l'organisme comparé. L'image de gauche montre une correspondance quasiment univoque entre le chromosome 1 de l'humain et celui du chimpanzé : l'ordre des gènes est conservé sur presque toute la longueur du chromosome, et leurs séquences sont identiques à 98%. On constate quelques réarrangements intra-chromosomiques (inversion de l'ordre de certains fragments), et quelques translocations extra-chromosomiques (deux fragments du chromosome 1 de l'humain se retrouvent sur les chromosomes 2 et 10 du chimpanzé, respectivement). La figure de droite montre que les réarrangements chromosomiques sont beaucoup plus nombreux quand on compare des organismes phylogénétiquement plus éloignés: les fragments de notre chromosome 1 se trouvent dispersés sur plus de 8 chromosomes du poulet.

On peut facilement étendre ce genre d'analyse en comparant tous les chromosomes entre plusieurs dizaines de génomes de métazoaires (animaux pluricellulaires)

actuellement disponibles, pour analyser les événements de réorganisations chromosomiques survenus au cours de l'évolution. On sait que de tels réarrangements contribuent à l'établissement des barrières interspécifiques. Un cas typique est celui des croisements entre un cheval et un âne : leur proximité génétique est encore suffisante pour qu'ils soient interféconds, mais leur descendance (mulets et baudets) est stérile, du fait de quelques réarrangements survenus au cours de la divergence entre ces espèces, qui empêchent l'appariement des chromosomes lors de la méiose.

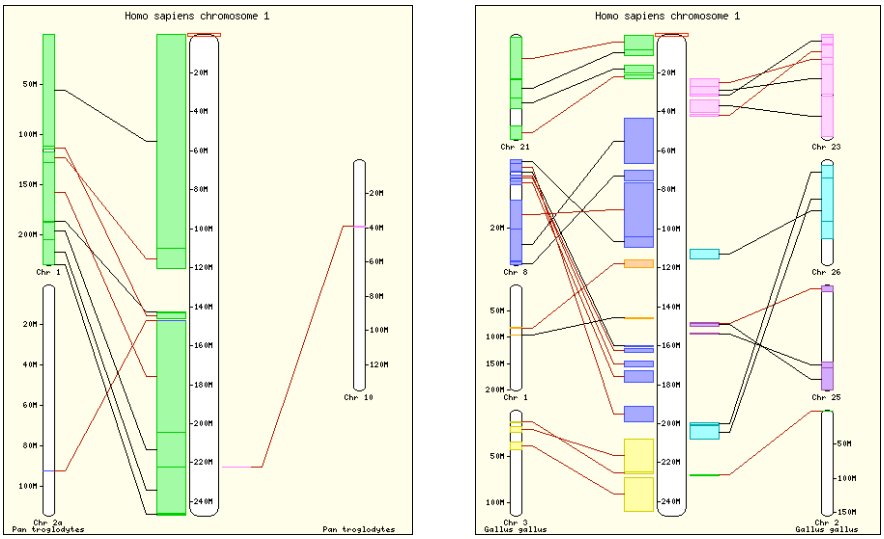

Figure 2. Les cartes « de synténie » indiquent la correspondance entre régions chromosomiques de deux organismes. Le chromosome 1 de l'humain a été comparé aux chromosomes de chimpanzé (image de gauche) et du poulet (image de droite). Ce graphique a été généré sur le site Web d'EnsEMBL (http://www.ensembl.org).

Les régions conservées révèlent les régions codantes et régulatrices

Nous pouvons également analyser les correspondances entre génomes à plus petite échelle, en nous focalisant sur un gène donné. La Figure 3 montre l'alignement des régions chromosomiques correspondant au gène *Pax6*, dont l'expression détermine la formation des yeux chez les animaux (voir l'article de René Rezsohazy dans ce volume). Chaque profil indique le pourcentage de nucléotides identiques entre la région *Pax6* d'un organisme donné et celle de l'humain. Au fur et à mesure que l'on s'éloigne de l'humain en terme de distance évolutive (de bas en haut), les pourcentages d'identité s'affaissent. Cependant, certaines petites régions montrent une plus grande tendance à la conservation. Les fragments les mieux conservés correspondent à des

parties codantes du gène (exons)[3], séparées par des régions transcrites non codantes (introns). De façon générale, les exons montrent un taux élevé de conservation du poisson à l'humain, par comparaison avec les régions intergéniques ou introniques.

On note également la présence de fragments non codants bien conservés dans les introns et dans les régions intergéniques. Ces conservations dans les régions non codantes correspondent généralement à des signaux de régulation, qui sont reconnus par des protéines spécialisées, les facteurs transcriptionnels, lesquels activent ou répriment la transcription du gène de façon extrêmement spécifique. C'est grâce à ces signaux de régulation que le gène *Pax6* s'exprime à un endroit précis et durant un moment précis durant le développement embryonnaire. Les cellules embryonnaires où s'exprime la protéine Pax6 formeront les yeux de l'animal. Il n'est pas étonnant que les régions codantes et les signaux de régulation du gène *Pax6* aient été conservés de façon aussi remarquable au cours de l'évolution : depuis des millions d'années, les individus ayant subi des mutations dans ces régions fonctionnelles naissent dépourvus d'yeux, et sont inexorablement éliminés par sélection naturelle[4].

L'exemple de Pax6 illustre la façon dont la génomique comparative permet de localiser les éléments fonctionnels essentiels, et d'analyser leur évolution au fil des espèces. Quoique le gène *Pax6* ait fait l'objet d'un grand nombre d'études depuis une dizaine d'années, il reste beaucoup à faire pour caractériser les éléments de régulation qui gouvernent son schéma[5] spatio-temporel d'expression durant le développement de l' embryon. Cependant, la génomique comparative permet déjà de délimiter des régions chromosomiques qui ont été conservées depuis des millions d'années en dépit des mutations qui bombardent en permanence les génomes. Cette conservation suggère que les mutations y ont été éliminées par la sélection, ce qui manifeste l'importance fonctionnelle de ces régions. La fonction précise de chaque région conservée pourra ensuite être analysée par les méthodes classiques de biologie moléculaire.

Une telle étude peut s'appliquer au cas par cas pour analyser les profils de conservation de chaque région chromosomique, et obtenir ainsi une vue détaillée de la divergence et de la conservation de chaque gène au fil de l'évolution.

---

[3] Notons également une forte conservation au début du premier exon et la fin du dernier exon, qui correspondent aux extrémités non traduites de l'ARN messager (régions marquées en jaune).

[4] Il existe des exceptions notables chez certains poissons cavernicoles (par exemple Astyanax), qui vivent dans l'obscurité complète, et pour lesquels la vision n'apporte pas d'avantage évolutif. Ces poissons sont dépourvus d'yeux, mais ont pu survivre grâce à un développement exceptionnel de la ligne latérale, un système mécano-récepteur qui leur permet de détecter les parois et les objets voisins.

[5] Le terme anglais précis est « pattern », dont la traduction littérale (« patron ») est ambiguë.

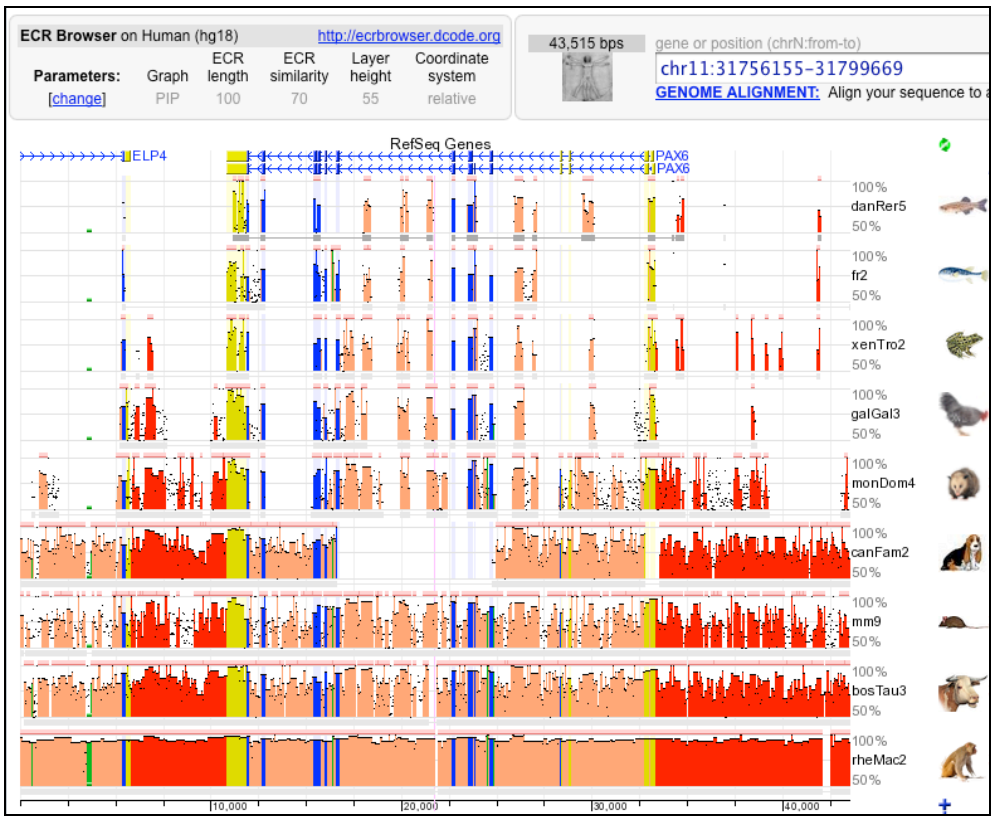

Figure 3. Comparaison des régions génomiques incluant le gène *Pax6* chez différents métazoaires. Les profils indiquent les pourcentages d'identité entre la région Pax6 de chaque organisme et celle de l'humain. Cette figure a été générée sur le site ECR Browser (http://ecrbrowser.dcode.org/). Note : la partie blanche dans le profil du génome canin correspond à une partie manquante dans la séquence génomique, suite à un problème technique lors du séquençage.

Des branches enchevêtrées dans les arbres phylogénétiques

L'unique figure du livre de Charles Darwin « *L'Origine des espèces* » (1859) montre un arbre de la vie schématique, illustrant son modèle général de divergence progressive entre groupes vivants (Figure 4a). La représentation arborescente reflète un mode de transmission essentiellement vertical du matériel génétique : les gènes sont transmis des parents aux enfants à chaque génération. Les mutations provoquent l'apparition de nouveaux caractères. L'isolement géographique favorise l'accumulation de différences entre populations, qui finissent par diverger jusqu'à former des espèces distinctes.

Au 20ème siècle, la génétique des bactéries a révélé des mécanismes permettant le transfert d'ADN entre bactéries appartenant à des espèces totalement différentes. De tels échanges, appelés « transferts horizontaux », ont tendance à brouiller les pistes

pour l'analyse des relations phylogénétiques entre organismes. Pour tenir compte des événements de transfert horizontal, Doolittle (1999) a proposé de représenter les relations entre espèces au moyen d'un arbre réticulé (Figure 4b).

Les échanges de matériel génétique sont encore plus fréquents chez les virus, qui ont une fâcheuse tendance à se recombiner pour former des espèces hybrides, dont la composition génétique est qualifiée de « mosaïque ». Gipsi Lima-Mendez a combiné les méthodes bioinformatiques de comparaisons de séquences et l'analyse des graphes pour proposer une représentation réticulée des relations entre bactériophages (des virus s'attaquant aux bactéries) (Figure 4c).

Ces trois exemples mettent en évidence la nécessité d'adapter nos modes de représentation pour tenir compte de la diversité des mécanismes sous-jacents à la transmission des caractères entre individus de la même espèce ou d'espèces distinctes.

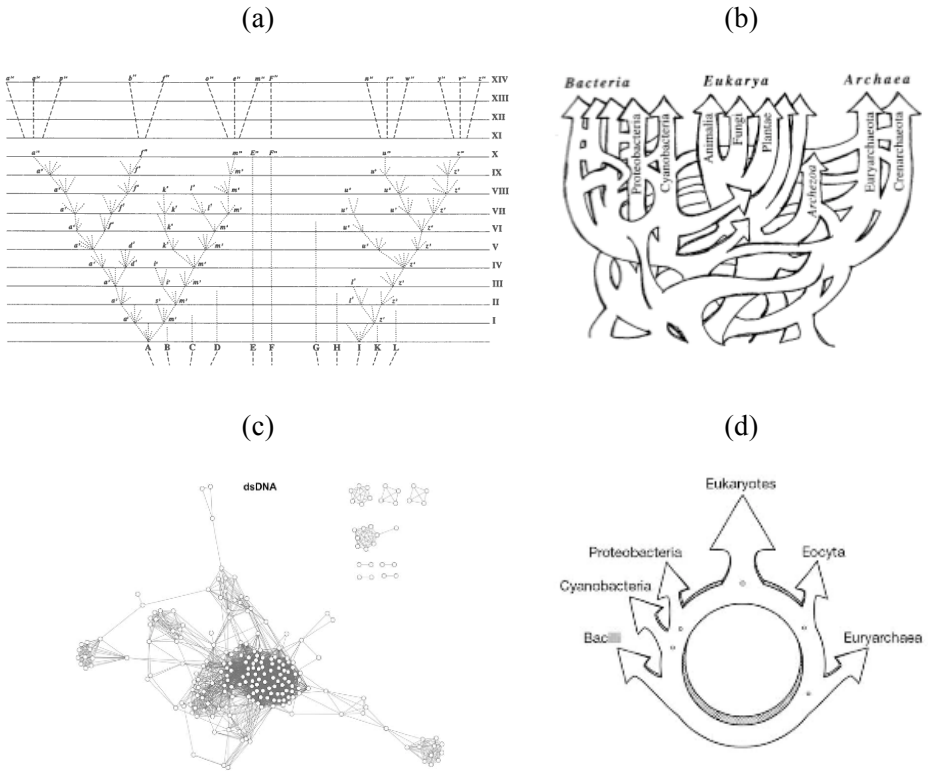

Figure 4. Des arbres, des réseaux et des anneaux. (a) L'arbre de la vie, unique graphique de *L'Origine des Espèces* (Darwin, 1859). (b) Représentation des transferts horizontaux sur un arbre réticulé (Doolittle, 1999). (c) une représentation réticulée (sous forme de réseau) souligne l'importance des échanges de matériel génétique entre espèces de bactériophages (Lima-Mendez, et al., 2008). (d) L'anneau de la vie proposé par Rivera et Lake (Rivera and Lake, 2004).

Les gènes fondamentaux

L'analyse phylogénique permet de remonter le temps bien au-delà de la divergence des vertébrés. Une comparaison systématique de tous les génomes disponibles a permis d'identifier plus ou moins 500 gènes qu'on retrouve chez tous les organismes. Ces gènes ont résisté aux outrages du temps depuis plus de 2 milliards d'années, raison pour laquelle Sean Carroll les qualifie d'immortels[6]. Ils sont impliqués dans des processus tellement fondamentaux (réplication de l'ADN, synthèse des protéines) que toute altération en a été impitoyablement éliminée au cours de l'évolution.

Pour des raisons statistiques, il est complètement invraisemblable que ces 500 gènes soient apparus de façon indépendante dans les 3 super-règnes (bactéries, archées, eucaryotes). Leur conservation renforce l'hypothèse d'un ancêtre commun à toutes les formes de vie terrestres actuellement connues[7].

A l'origine des eucaryotes : l'anneau de la vie

La génomique comparative nous éclaire également quant aux relations phylogénétiques entre les trois super-règnes. Lorsque les premières espèces d'archées ont été découvertes, on les a d'abord traitées comme un sous-groupe des bactéries (elles étaient d'ailleurs initialement appelées « archéobactéries »). La génétique moléculaire a ensuite révélé que, par certains aspects, les archées ressemblent plus aux eucaryotes qu'aux bactéries. Les biologistes ont donc proposé un scénario évolutif basé sur une séparation initiale entre bactéries et archées, suivie d'une séparation entre archées et eucaryotes.

Cependant, une analyse exhaustive des gènes d'eucaryotes, d'eubactéries et d'archées révèle que pour une partie des gènes eucaryotes, l'homologue le plus proche se trouve chez les bactéries, tandis que pour une autre partie il se trouve chez les archées. Pour résoudre cet apparent paradoxe, Rivera et Lake ont proposé un modèle original selon lequel les premières cellules eucaryotes résulteraient de la fusion entre une bactérie et une archée (Rivera and Lake, 2004). Sur base de ce modèle, ils proposent de représenter les relations entre super-règnes sous la forme d'un « anneau de la vie » (Figure 4d).

Les duplications de gènes à l'origine de l'innovation

Les duplications jouent un rôle important dans l'apparition de nouvelles fonctions géniques. L'analyse des génomes montre des traces de nombreuses duplications d'ampleur variable : courtes régions englobant un ou plusieurs gènes, chromosomes entiers, voire génomes complets (avec formation d'individus polyploïdes, c'est-à-dire possédant plusieurs copies de chaque chromosome).

---

[6] Sean Carroll entend par là qu'ils ont une longévité exceptionnelle, puisqu'ils ont été conservés pendant plus de 2 milliards d'années. Cette « immortalité » ne signifie pas qu'ils survivront jusqu'à la fin des temps. Le terme « fondamentaux » indique ici qu'ils participent au fondement, aux bases de la vie.

[7] On appelle *LUCA* (*Last Universal Common Ancestor*) l'ancêtre commun le plus récent entre toutes les formes de vie connues.

Le premier effet de la duplication d'un gène est de provoquer une redondance fonctionnelle, qui n'est généralement ni nuisible ni bénéfique à l'organisme. L'une des deux copies peut dès lors subir des mutations, et progressivement diverger de l'original sans que l'organisme en soit affecté, puisque l'autre copie continue à assurer la fonction initiale. Occasionnellement, une telle mutation peut modifier la fonction du gène de façon bénéfique à l'organisme. Je reprends ici deux exemples frappants présentés par Sean Carroll dans *The making of the fittest* (2006).

Chez les animaux, la perception de la lumière est assurée par un pigment (la rhodopsine) composé d'une protéine (opsine) et d'un groupement prosthétique (rétilène). La capacité à discerner les couleurs résulte de la présence, dans notre génome, de plusieurs gènes codant des opsines sensibles à différentes longueurs d'ondes. La plupart des animaux ne disposent que de deux opsines, qui leur confèrent une vision dichromatique. La vision trichromatique est apparue chez les primates de l'ancien monde, du fait de la duplication de l'opsine sensible au vert. Suite à cette duplication, il a suffi de quelques mutations ponctuelles d'une des deux copies pour provoquer un décalage du spectre de sensibilité vers le rouge.

Les duplications de gènes peuvent se produire de façon répétée, et générer de grandes familles de gènes homologues[8]. L'exemple le plus frappant est celui des récepteurs olfactifs (Niimura and Nei, 2003; Niimura and Nei, 2007): les génomes des mammifères contiennent plusieurs centaines de gènes codant des récepteurs olfactifs. Chez la souris, on n'en compte pas moins de 1400, soit 5% des gènes de tout le génome. Chacun de ces gènes code une protéine capable de se lier à des molécules spécifiques. La finesse de l'odorat des mammifères résulte de l'accumulation de ces récepteurs qui leur ont permis, au cours de l'évolution, de détecter un nombre de plus en plus diversifié de substances chimiques dans leur environnement (l'air, la nourriture).

Ces exemples montrent que les combinaisons de duplications/divergence/sélection constituent un mécanisme fertile pour enrichir le répertoire fonctionnel des gènes au cours de l'évolution.

Duplications de génomes entiers

Occasionnellement, ces mécanismes prennent place à l'échelle d'un génome entier. C'est le cas d'un grand nombre de variétés de plantes, qui se distinguent par la présence de copies multiples de tous les chromosomes. La polyploïdie résulte d'accidents génétiques survenus lors de la formation des cellules germinales, et elle provoque souvent une augmentation de la taille des cellules et de la plante. La sélection exercée depuis 10.000 ans par les agriculteurs a apparemment favorisé les lignées polyploïdes pour plusieurs de nos variétés cultivées.

[8] Le terme *homologue* s'applique à des gènes qui présentent des séquences similaires du fait d'une origine évolutive commune. L'origine commune peut remonter à une divergence entre espèces (on parle alors d'*orthologie*: l'hémoglobine alpha humaine est *orthologue* à l'hémoglobine alpha canine) ou à une duplication de gène (on parle alors de *paralogie*: l'hémoglobine alpha humaine est *paralogue* à l'hémoglobine beta humaine).

Sur base de séquences génomiques, on peut détecter des duplications de génomes, soit par comparaison de plusieurs génomes apparentés, soit en analysant le nombre de copies de chaque gène dans un seul génome. En comparant les génomes entiers de 2 levures (*Kluyveromyces lactis* et la levure du boulanger *Saccharomyces cerevisiae*), Kellis et coll. ont montré qu'un événement de duplication du génome entier avait eu lieu au cours de l'histoire des levures (Kellis, et al., 2004). On a également montré que le génome de la paramécie (protozoaire) avait subi 3 cycles de duplication complète (Aury, et al., 2006), de sorte qu'on retrouve de multiples copies de chaque gène (3 duplications successives donnent 8 copies de gènes).

Les duplications de génomes entiers ouvrent bien entendu la voie à une diversification massive des fonctions présentes dans un génome, puisque tous les gènes se retrouvent dupliqués : pour chaque gène, tant que l'une des deux copies assure la fonction initiale, l'autre copie pourra subir des mutations dont la plupart n'auront pas d'effet sensible sur l'aptitude de l'organisme. Occasionnellement, certaines de ces mutations pourront susciter l'apparition de nouvelles fonctions.

Les gènes fossiles

On détecte dans les génomes des morceaux de séquence ressemblant fortement à des gènes connus, mais qui ne peuvent en aucun cas produire de protéine fonctionnelle, car ils sont truffés de codons stop. Lafontaine et coll. ont détecté 149 régions de ce type dans le génome de la levure (Lafontaine, et al., 2004). Ils interprètent ces régions comme des « reliques » de gènes anciennement fonctionnels.

Sean Carroll (2006) présente une série de cas de ces gènes qu'il qualifie de « fossiles »: gènes qui ont perdu leur activité, mais dont on retrouve des traces dans les génomes. Selon Carroll, cette fossilisation succède à un relâchement de la pression sélective.

Tous les gènes sont en permanence soumis au bombardement des mutations, mais les mutations qui affectent la fonction d'un gène sont généralement éliminées par la sélection naturelle (sélection « purificatrice »). Si, pour une raison ou une autre, un gène cesse d'être indispensable pour un organisme donné dans un environnement donné, cette pression sélective est amoindrie, et les mutations s'accumulent.

Un superbe exemple est celui des poissons des glaces (famille des Channichthyidae) vivant dans l'Arctique, dans des eaux dont la température varie de 4°C à -2°C. Les Channichthyidae survivent à ces températures grâce à une série de transformations physiologiques, qui permettent d'éviter à leur sang de se congeler. Un premier mécanisme repose sur la présence, dans leur sang, des protéines « antigels », composées de motifs répétitifs. Une autre transformation est l'absence totale de globules rouges, qui a pour effet de diminuer la viscosité du sang, et facilite donc sa circulation à basses températures. La respiration de ce poisson est assurée par diffusion de l'oxygène dans le liquide sanguin, sans qu'aucune hémoglobine ne soit synthétisée. Cependant, on trouve dans leur génome des gènes fossilisés pour les deux chaînes de l'hémoglobine.

Un exemple de fossilisation massive est celui des récepteurs olfactifs dont nous avons parlé plus haut. Parallèlement aux nombreuses duplications qui ont provoqué

l'expansion de cette famille de gènes, il arrive relativement fréquemment que ces gènes soient affectés par l'une ou l'autre mutation qui annihile leur fonction. L'effet de ces mutations est cependant relativement marginal: l'organisme perd la capacité à détecter l'une ou l'autre odeur, mais son répertoire olfactif contient toujours quelques centaines de gènes fonctionnels. Niimura et Nei (2007) estiment que, chez l'humain, cette « fossilisation » touche plus de 50% des 800 gènes codant des récepteurs olfactifs, alors que chez la souris, elle touche 25% des 1400 récepteurs identifiés dans le génome.

Avec l'augmentation du nombre de génomes disponibles, on trouve de plus en plus d'exemples de fossilisation de gènes, et en comparant les génomes d'espèces plus ou moins éloignées, on peut retracer l'histoire des duplications, délétions et inactivations des gènes.

La présence même de gènes fossiles illustre une fois de plus le fait que l'organisation des organismes vivants ne relève pas d'un « plan » conçu par un ingénieur (une intelligence supérieure) pour répondre de façon optimale à un dessein prédéfini. Au contraire, nous observons ici les traces d'un « bricolage » évolutif, pour reprendre l'expression de François Jacob (Jacob, 1977).

Le rêve du génome humain[9]

Le projet génome humain a suscité un engouement médiatique, essentiellement motivé par les perspectives médicales que faisaient miroiter ses promoteurs. Nous assistons aujourd'hui à l'avènement de la génomique personnelle, dont les champs d'application vont de la médecine personnalisée à la police scientifique en passant par le traçage des origines ethniques de chaque individu. À moyen terme, il ne fait aucun doute que la génomique contribuera aux progrès de la médecine (médecine préventive, médecine personnalisée, thérapie génique), et par là même au bien-être de l'humanité. Cependant, dans une société gouvernée par un pouvoir totalitaire, ou dans un monde où la loi du marché prévaudrait sur les valeurs éthiques, les risques de dérive sont nombreux. Nous avons développé ailleurs les enjeux éthiques liés aux développements de la génomique personnelle (van Helden, 2008). Nous ne citerons qu'un exemple tout récent: en novembre 2008, la revue Nature publie une étude menée sur plus de 3000 individus, montrant qu'il est possible, sur simple base de profils génomiques, de retracer l'origine géographique d'un Européen dans un rayon de 500 km (Novembre, et al., 2008). De telles études présentent un grand intérêt notamment pour retracer l'histoire des populations européennes, mais on imagine facilement les risques que présentent ces développements technologiques dans un contexte de violence ethnique comme en ont connu les Balkans il y a peu de temps (Detours, 2008).

[9] Le titre de cette section se réfère à un article publié par Richard Lewontin dans les premières années du projet génome humain (Lewontin, R. (1992). The dream of the human genome: doubts about the Human Genome Project. *New York Rev Books* **39,** 31-40.). Lewontin y exprime ses doutes concernant l'apport scientifique à attendre de ce projet, et ses craintes concernant certaines applications de la génomique.

Conclusion

L'analyse des génomes a pris naissance il y a une dizaine d'années, et connaît un essor croissant. Au stade actuel, le premier enjeu reste de développer les approches qui nous permettront de comprendre la façon dont ces génomes fonctionnent. Il est clair que nous en sommes encore aux premiers balbutiements, et il ne fait aucun doute que le décryptage des génomes occupera encore certainement quelques générations de biologistes et bioinformaticiens.

L'espace est beaucoup trop restreint pour pouvoir brosser un tableau général des questions et apports de la génomique. Je n'ai développé ici que quelques exemples qui montrent la façon dont l'analyse des génomes nous permet de dévoiler certains événements passés, et de comprendre les mécanismes qui président à l'évolution. Le lecteur intéressé trouvera une mine d'informations et de superbes exemples dans l'excellent ouvrage *The Making of the Fittest* (Carroll, 2006).

Remerciements

Je remercie Annick Stevens, Ariane Ramaekers et Jean Vandenhaute pour leurs corrections, commentaires et suggestions concernant ce manuscrit.

Sites Web de référence

Ce chapitre fait la synthèse d'un exposé de l'Ecole d'Eté « *Penser l'Evolution* », qui s'est tenue en août 2008 à l'Université Libre de Bruxelles. Les supports graphiques et enregistrements sonores de ces trois journées sont disponibles sur demande en contactant les organisateurs de Penser la Science (http://www.ulb.ac.be/penser-la-science/).

EnsEMBL	http://www.ensembl.org	Site de référence pour la visualisation et l'analyse des génomes d'organismes pluricellulaires.
NCBI	http://www.ncbi.nlm.nih.gov/	Le National Center for Biotechnology Information (Etats-Unis) donne accès à toutes les séquences publiées, et contient une importante collection de génomes complets.
ECR Browser	http://ecrbrowser.dcode.org/	Un outil très pratique pour visualiser les génomes.

Références bibliographiques

Aury, J. M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., Segurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Camara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A. M., Kissmehl, R., Klotz, C., Koll, F., Le Mouel, A., Lepere, G., Malinsky, S., Nowacki, M., Nowak, J. K., Plattner, H., Poulain, J.,

Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Betermier, M., Weissenbach, J., Scarpelli, C., Schachter, V., Sperling, L., Meyer, E., Cohen, J. and Wincker, P. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444,** 171-8.

Carroll, S. B. (2006). *The Making of the Fittest. DNA and the Ultimate Forensic Record of Evolution*. Norton.

Darwin, C. (1859). *L'origine des espèces*, 1992 edn. Flammarion.

Detours, V. (2008). Editorial comment should accompany hot papers online. *Nature* **455,** 861.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* **284,** 2124-9.

Fox-Keller, E. (2003). *Le siècle du gène*. Gallimar.

Jacob, F. (1977). Evolution and tinkering. *Science* **196,** 1161-6.

Kellis, M., Birren, B. W. and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428,** 617-24.

Lafontaine, I., Fischer, G., Talla, E. and Dujon, B. (2004). Gene relics in the genome of the yeast Saccharomyces cerevisiae. *Gene* **335,** 1-17.

Lewontin, R. (1992). The dream of the human genome: doubts about the Human Genome Project. *New York Rev Books* **39,** 31-40.

Lima-Mendez, G., Van Helden, J., Toussaint, A. and Leplae, R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol* **25,** 762-77.

Niimura, Y. and Nei, M. (2003). Evolution of olfactory receptor genes in the human genome. *Proc Natl Acad Sci U S A* **100,** 12235-40.

Niimura, Y. and Nei, M. (2007). Extensive gains and losses of olfactory receptor genes in Mammalian evolution. *PLoS ONE* **2,** e708.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M. and Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*.

Rivera, M. C. and Lake, J. A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431,** 152-5.

van Helden, J. (2008). Qui lira notre avenir dans nos gènes ? *Lettre de l'Académie Royale de Belgique* **sous presse**.