Préface

Nos sociétés modernes sont trop préoccupées par leurs performances immédiates pour imaginer qu'un monde peut exister où les coûts et les efforts déployés pour atteindre un résultat ne seraient pas rationnellement minimisés. C'est pourtant l'image que nous offre le monde vivant dès que l'on prend soin de l'étudier attentivement. La remarquable adaptation des différents organismes à leurs conditions de vie repose sur des génomes qui sont loin d'être les produits de ce qui nous semblerait une ingénierie rationnelle. Il n'existe pas de génomes minimums. Tous sont trop grands par rapport au nombre de gènes que l'on estime nécessaires à produire l'organisme qui les porte. Souvent beaucoup trop grand. Le nôtre apparait cinquante fois trop grand. Et tous contiennent des séquences répétées, à l'identique ou presque, jusqu'à de nombreuses fois dans le même génome alors que les combinaisons aléatoires des quatre nucléotides rendent ce phénomène extrêmement improbable, voire pratiquement impossible.

Cet état de fait avait été observé dès le milieu du XX^{e} siècle, bien avant l'émergence de la génomique, par l'étude des cinétiques de renaturation des molécules d'ADN. L'excès d'ADN et l'abondance de séquences répétées restaient une énigme que certains avaient tendance à balayer rapidement en parlant d'ADN poubelle (*junk DNA*) et en continuant à étudier seulement ce qu'ils connaissaient déjà ! Le séquençage des génomes allait résoudre cette énigme en montrant à quel point nos connaissances antérieures étaient incomplètes. Une nouvelle vision de l'organisation des génomes et de leur fonctionnement s'offre maintenant à nous dans lesquelles la dynamique temporelle se conjugue avec le présent. Car tous les génomes ne sont que des copies imparfaites des génomes qui les ont précédés, pas des constructions neuves. Dès lors, les traces du passé se mêlent aux évènements présents et, ensemble seulement, ils offrent les bases de l'avenir.

Fonction et évolution des séquences répétées dans les génomes,
coordonné par Guy-Franck RICHARD. © ISTE Editions 2024.

Comme l'illustre remarquablement cet ouvrage, les répétitions que l'on trouve dans les génomes peuvent provenir d'accidents évolutifs importants comme les duplications totales de génomes qui, phénomène remarquable, tendent à coïncider dans le temps avec les transitions entre les grandes époques géologiques. Mais elles peuvent aussi provenir d'interactions répétées avec des éléments infectieux – types viraux - qui finissent par s'intégrer dans les chromosomes et être transmis à la descendance. Il y a donc des causes endogènes et des causes exogènes à l'existence de séquences répétées dans les génomes. À leur tour, les répétitions peuvent être la base de la formation d'éléments fonctionnels des chromosomes comme les centromères, les télomères ou les quadruplex à guanine. Les variations des nombres de copies répétées de séquences longues peuvent jouer un rôle critique dans les phénotypes et dans l'adaptation des organismes. De même, les instabilités de répétitions de séquences courtes permettent de différencier facilement les individus d'une même population, mais peuvent parfois conduire à des syndromes très graves. Enfin, on ne peut pas ignorer les différents éléments mobiles, sortes de machines moléculaires spécialisées, présents en nombre variable dans les différents génomes. Ils avaient été repérés dès le milieu du XX^{e} siècle par leurs effets génétiques – ils sont mutagènes –, mais on dispose maintenant d'une vision beaucoup plus large de leur diversité et des conséquences, parfois considérables, de leur activité.

Si nos connaissances sur les séquences répétées des génomes n'ont progressé que tardivement, c'est en partie à cause des difficultés techniques de leur séquençage. Jusqu'à l'arrivée récente des nouvelles technologies permettant des lectures longues, il était très difficile d'assembler correctement les lectures de séquences répétées et, de fait, de très nombreuses séquences génomiques dites complètes ne l'étaient pas. À titre d'exemple, environ 8 % du génome humain, fait de séquences hautement répétées, sont restés inconnus pendant deux décennies jusqu'à l'application de technologies spéciales cette année. De même, l'étude des variations de nombre de copies dues aux duplications segmentaires, longtemps sous-estimées, ne fait que débuter. Et il ne faut pas oublier que notre exploration du monde vivant est non seulement loin d'être complète, mais est très biaisée en faveur des groupes d'organismes déjà bien connus. Il faut donc s'attendre à de nouvelles découvertes, voire surprises, dans l'étude de cette partie des génomes trop longtemps ignorée qui nous démontre combien le véritable succès à long terme diffère de l'illusion des performances immédiates.

Gif-sur-Yvette

Bernard DUJON
Professeur émérite à Sorbonne Université et à l'Institut Pasteur
Membre de l'Institut de France (Académie des Sciences)

Introduction

Des génomes répétés

Guy-Franck RICHARD
Institut Pasteur, CNRS, Université Paris Cité, Sorbonne Université, Paris, France

DÉFINITION. Génome [ʒenom] *n.m. Biol.* Ensemble des caractères héréditaires d'un être vivant, composé pour une petite partie de gènes assurant une fonction dans l'organisme et pour la plus grande partie de séquences répétées dont on ignore si elles ont une fonction.

En forçant à peine le trait, telle pourrait être une définition moderne du mot « génome », à l'aune des connaissances apportées depuis trois décennies par le séquençage du contenu en ADN des êtres vivants, en particulier des organismes eucaryotes, plus complexes que ceux de leurs ancêtres bactériens et archées. Les biologistes savaient depuis les années 1960, bien avant l'invention des premières méthodes de séquençage de l'ADN, que le contenu des génomes était compliqué à appréhender. Les expériences de dénaturation-renaturation avaient fait apparaître que la vitesse de renaturation de la double hélice était proportionnelle à sa concentration. Le paramètre $C_o t$ était la valeur à laquelle la renaturation de la moitié de l'ADN génomique était complète, dans des conditions contrôlées. Chaque organisme pouvait alors être défini par la valeur $C_o t$ de son génome. En essayant d'établir les valeurs $C_o t$ de génomes d'organismes les plus simples – phages ou bactéries – ou plus complexes comme ceux des vertébrés, il apparaissait que ces derniers contenaient trois types de séquences aux valeurs $C_o t$ très différentes (figure I.1).

On peut ainsi montrer que le génome de la souris, par exemple, est composé pour 70 % de séquences uniques à la renaturation lente, pour 20 % de séquences modérément répétées présentes de 1 000 à 100 000 copies par génome et pour 10 % de séquences hautement répétées représentant au moins un million de copies par génome et montrant une renaturation rapide (Britten et Kohne 1968). Cette approche, basée sur les propriétés physico-chimiques de l'ADN, sous-estimait légèrement la quantité de séquences répétées,

Fonction et évolution des séquences répétées dans les génomes,
coordonné par Guy-Franck RICHARD.

car leur vitesse de renaturation dépend de l'identité entre ces séquences, les séquences divergentes (comme les LTR) se renaturant plus lentement que les séquences identiques. De nos jours, les courbes de C_0t sont encore parfois utilisées pour séparer la fraction hautement répétitive d'un génome de sa fraction unique, dans le but de séquencer l'ADN spécifique de l'une ou l'autre fraction (Peterson *et al.* 2008).

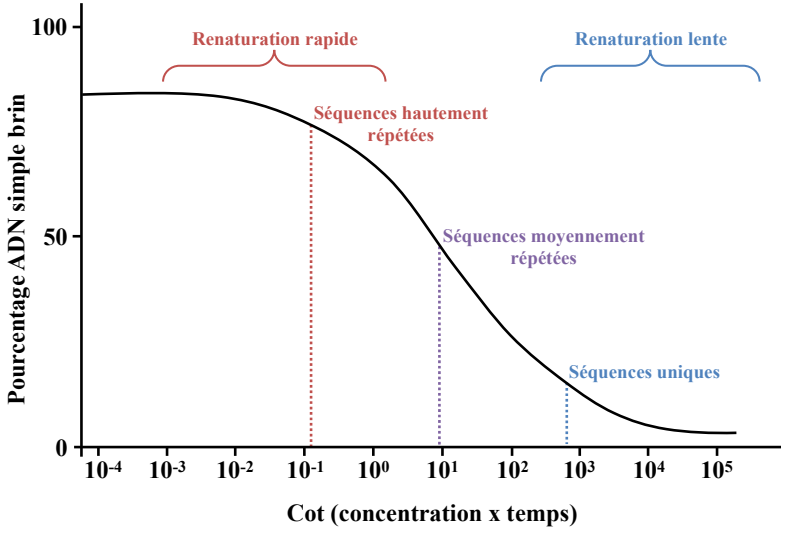

Figure I.1. *Exemple de courbe de C_0t*

I.1. Le paradoxe de la « valeur C »

À partir du moment où il fut prouvé que l'ADN était le support de l'hérédité, et contenait théoriquement tous les gènes nécessaires au développement d'un être vivant, il paraissait logique que les organismes les plus sophistiqués devaient contenir plus de gènes et donc plus d'ADN (la « valeur C ») dans leur génome pour coder ces gènes. Cette idée allait être remise en question dans les années 1950 avec la découverte que les noyaux de certains amphibiens et poissons contenaient 20 fois plus d'ADN que les noyaux des mammifères. Ces derniers présentant une complexité développementale bien plus grande, cela apparut comme tout à fait paradoxal, et fut même utilisé comme argument par les contempteurs de l'ADN comme seul support de l'hérédité (Thomas 1971). Ce « paradoxe de la valeur C » ne trouva finalement son explication que des décennies plus tard, lorsque les premiers génomes furent séquencés. On sait maintenant que le nombre de gènes d'un organisme n'a que peu à voir avec sa taille ou avec son niveau de complexité. Le génome de la levure de boulangerie contient environ 6 000 gènes, celui de la mouche du vinaigre environ 14 000 et le génome humain (ou ceux de ses très proches cousins les grands singes) se contente de 20 000 gènes, avec lesquels il gère un niveau de complexité

développementale et comportementale très sophistiqué. Mais que penser alors de la paramécie avec ses 40 000 gènes, deux fois plus que le génome humain ? Ou de *Trichomonas vaginalis*, un parasite du tractus génital, et de ses 60 000 gènes ? Ou bien du blé et de ses 124 000 gènes, plus de six fois le nombre de nos gènes ? La prétendue complexité ne pouvait clairement se mesurer au nombre de gènes d'un organisme. Les études de génomique comparative[1] ont montré que ce nombre élevé de gènes dans certains organismes cache en fait des évènements ancestraux de duplication partielle ou totale de génomes, suivis de pertes de gènes plus ou moins importantes (Wolfe et Shields 1997 ; Jaillon *et al.* 2004). Ces évènements participent activement à la redondance génétique et leur mise en évidence ainsi que les mécanismes qui les sous-tendent seront les thèmes abordés dans le chapitre 1.

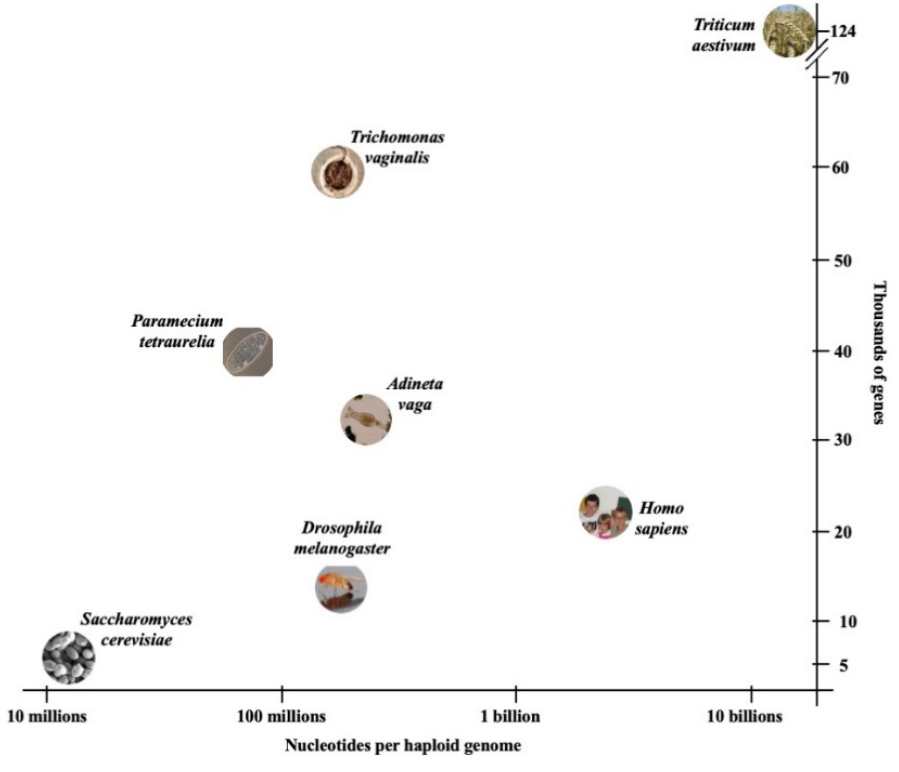

Figure I.2. *Comparaison de la taille de quelques génomes et du nombre de gènes qu'ils contiennent*

1. La génomique comparative est le domaine de la génomique qui s'intéresse à la comparaison des génomes entiers entre eux et non plus seulement des gènes. Des outils d'analyse ont été spécifiquement développés afin de comparer l'organisation, la structure, la synténie (ordre des gènes le long d'un chromosome) des génomes, considérés comme des objets d'étude à part entière.

Si la complexité d'un organisme n'a rien à voir avec le nombre de gènes qu'il contient, il en est de même de la quantité d'ADN. Le génome humain avec un peu plus de trois fois le nombre de gènes de la levure de bière contient 200 fois plus d'ADN. Le génome d'un rotifère – petit animal de quelques millimètres vivant dans les eaux douces – contient 50 % de gènes en plus que le génome humain dans douze fois moins d'ADN (figure I.2).

La séquence génomique de tous ces organismes a montré que certains d'entre eux avaient évolué vers un génome très compact, à la densité élevée de gènes, alors que d'autres contenaient une multitude de séquences d'ADN répété dont la fonction n'apparaissait pas évidente à première vue et que certains auteurs n'hésitèrent pas à qualifier d'ADN « poubelle » (Ohno 1972).

I.2. Le recyclage de l'ADN poubelle

Environ 2 % du génome humain est traduit en protéines. Même en ajoutant les gènes non traduits (ARNt, ADNt, siARN, snARN, etc.), le pourcentage d'ADN « utile » augmente à peine. À quoi servent donc les 98 % d'ADN de notre génome qui n'ont – apparemment – pas de fonction ? Une des réponses envisageables est qu'ils n'en ont aucune. Le consortium emmené par Jeff Boeke, professeur de génétique de la Johns Hopkins University à Baltimore, s'était donné pour but de créer la première levure synthétique, à partir d'oligonucléotides de synthèse. La levure de bière, *Saccharomyces cerevisiae*, est un organisme eucaryote dont le génome contient 12,5 millions de nucléotides répartis sur 16 chromosomes. Les chromosomes synthétiques furent reconstruits un à un à partir de séquences de 70 nucléotides assemblées en blocs de 750 paires de bases, eux-mêmes assemblés en mégablocs de 2 à 4 kilobases, réintroduits les uns après les autres, de façon hiérarchisée, dans le génome de la levure en remplacement des séquences naturelles (Muller et Koszul 2015). Lors de la conception des chromosomes synthétiques, il fut décidé de retirer toutes les séquences répétées du génome. Tous les ADN codant les ARNt furent regroupés sur un seul chromosome circulaire, construit à façon pour les contenir. Les rétrotransposons, les microsatellites, les minisatellites et autres éléments répétés non indispensables à la vie furent effacés de la nouvelle séquence. Ces chromosomes synthétiques débarrassés de leur ADN poubelle sont parfaitement capables de faire vivre la levure qui les contient, sans aucun défaut phénotypique apparent, du moins dans les conditions de croissance en laboratoire (Dymond *et al.* 2011 ; Annaluru *et al.* 2014). On pourrait conclure des résultats de ce projet que l'ADN poubelle ne sert à rien. Ce serait une erreur.

Le génome humain de référence contient environ 443 000 éléments résiduels d'invasions rétrovirales passées, couvrant 8,3 % de la séquence totale (*International Human*

Genome Sequencing Consortium 2001). Ces cicatrices rétrovirales sont les restes d'invasions successives, depuis une centaine de millions d'années, de nos ancêtres mammifères par des éléments exogènes ayant laissé la trace de leur passage sous la forme de LTR[2]. Ces restes rétroviraux font donc partie de notre ADN poubelle. Néanmoins, comme nous allons le voir, leur présence dans notre génome témoigne de leur lointain mais indispensable rôle dans l'existence de notre lignée. Les mammifères thériens, c'est-à-dire ceux possédant un utérus au sein duquel se développe l'œuf fécondé, sont classés en deux groupes. Les euthériens (ou placentaires) comme l'homme ou la souris possèdent un placenta très élaboré connectant la paroi de l'utérus à l'embryon et permettant à celui-ci de se développer en toute sécurité pendant toute la période de la gestation. Les marsupiaux (kangourous et koalas) ne possèdent pas de placenta, et le développement de leurs petits se fait principalement en dehors de l'utérus. Le séquençage des génomes a montré que les deux gènes humains spécifiquement exprimés dans le placenta, *syncytin-1* et *syncytin-2*, dérivaient d'un gène codant une protéine virale ancestrale ayant infecté la lignée des primates il y a 25-40 millions d'années. De façon remarquable, le génome de la souris, autre mammifère placentaire, contient également deux gènes viraux ayant la même fonction que les gènes humains, mais dérivant d'une infection virale légèrement plus récente que celle de la lignée humaine. Le placenta a donc été inventé deux fois, indépendamment, dans deux lignées de mammifères, par capture de gènes d'origine rétrovirale (Dupressoir *et al.* 2009). Un autre exemple est encore plus frappant. La reproduction sexuée a été inventée à l'origine du monde eucaryote. À partir des premières cellules eucaryotes primitives s'est développé un système de syngamie[3] permettant de fusionner les noyaux de deux cellules haploïdes pour donner naissance à une cellule diploïde. La protéine responsable de la fusion des gamètes mâle et femelle est la même chez les plantes et les animaux, c'est le produit du gène *HAP2*. Cette protéine est d'origine virale et permet à l'enveloppe d'un virus de fusionner avec la membrane plasmique des cellules de son hôte (Fédry *et al.* 2017). Ainsi, un gène essentiel à la reproduction sexuée a été capturé depuis un virus par le génome des toutes premières cellules eucaryotes il y a environ 1,5 milliard d'années.

D'autres exemples existent de capture d'un morceau d'élément transposable, créant ainsi un nouveau gène, une nouvelle fonction. L'ADN poubelle est donc régulièrement recyclé au cours de l'évolution afin d'apporter diversité et nouveauté. Comme le disait François Jacob il y a déjà plus de 40 ans, l'évolution « bricole », elle fait du neuf avec du vieux, réutilisant des bouts de gènes, les coupant, les collant, les fusionnant avec d'autres afin de créer de la nouveauté (Jacob 1977). Ce qui apparaît aujourd'hui aux généticiens du XXIe siècle comme de l'ADN poubelle a peut-être servi par le passé – ou servira dans

2. LTR (*long terminal repeat*) : séquences répétées caractéristiques trouvées aux sites d'insertion des rétrovirus.

3. Syngamie : fusion des noyaux de deux cellules de type sexuel opposé.

le futur – à créer de la diversité. La formidable réussite du monde eucaryote à envahir toutes les niches écologiques sous tous les climats et toutes les latitudes provient en partie de l'extraordinaire flexibilité de son génome et de sa capacité à accumuler des éléments génétiques en apparence inutiles, mais qui sur le long terme vont être recyclés pour créer de la nouveauté et permettre l'apparition de nouvelles espèces vivantes.

I.3. Les différents types de répétitions

Il y a souvent plusieurs façons de classer des éléments génétiques. Certains auteurs ont choisi de distinguer les éléments répétés dispersés par opposition aux éléments répétés en tandem, ces derniers étant répétés au moins deux fois l'un derrière l'autre au même *locus* génétique, à l'inverse des premiers qui sont répétés à des loci différents (Richard *et al.* 2008).

Mais certaines répétitions dispersées sont tellement nombreuses dans les génomes qu'elles apparaissent comme étant répétées en tandem. C'est le cas des séquences *Alu* chez l'homme, qui sont fréquemment trouvées groupées dans des introns ou des séquences intergéniques.

On pourrait aussi distinguer les séquences répétées d'origine exogène, c'est-à-dire provenant d'un autre organisme que la cellule où on les observe, des séquences répétées d'origine endogène, fabriquées par la cellule où on les observe. Les éléments transposables feraient partie de la première catégorie, ayant envahi les génomes des lignées eucaryotes (ou procaryotes), alors que les différents ADN satellites appartiendraient à la deuxième, étant fabriqués par des processus moléculaires propres aux génomes qui les contiennent. Mais d'autres problèmes se posent alors.

On sait par exemple que les éléments *Alu*, rétrotransposons inactifs pouvant être mobilisés en *trans* par la machinerie d'autres rétroéléments, ont une origine endogène. Ils résultent en effet de la duplication de l'ARN non codant 7SL, impliqué dans la synthèse des protéines excrétées. Cette duplication, antérieure à la radiation des mammifères, aboutit à la fusion de deux monomères de 130 nucléotides dérivés de l'ARN 7SL, séparés par une courte région riche en adénines (Ullu et Tschudi 1984). Atteindre une classification cohérente des séquences répétées s'avère donc une tâche compliquée, particulièrement dans les génomes de plantes et d'animaux évolués au sein desquels elles sont pléthoriques, tant en structure qu'en nombre.

Nous avons donc essayé dans la suite de cet ouvrage de présenter les éléments répétés en rapport avec leur rôle (prouvé ou supposé) dans les génomes plus qu'en fonction de leur structure ou de leur origine supposée (figure I.3).

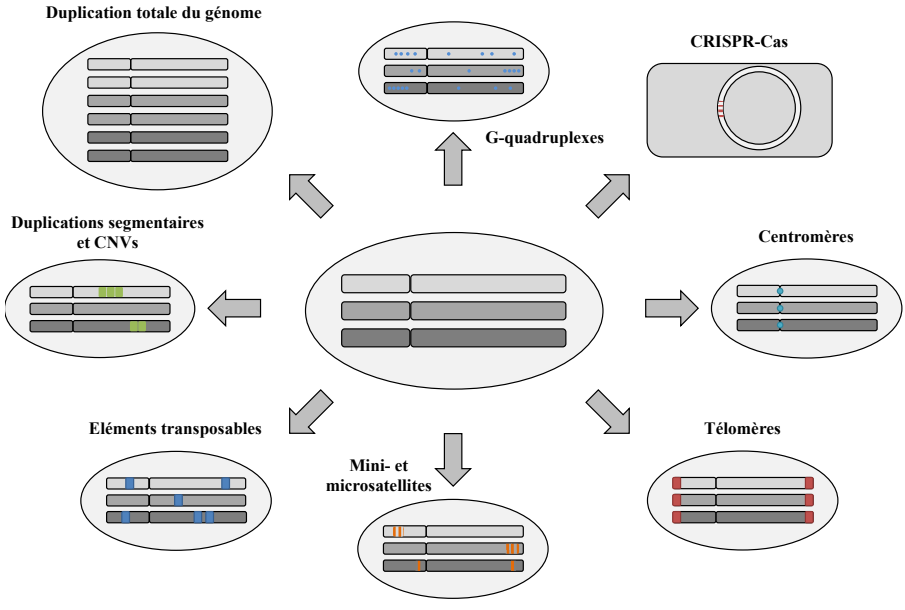

Figure I.3. *Les différents types d'ADN répété*

Après les duplications totales ou partielles de génomes vues au chapitre 1, les duplications de larges segments d'ADN, parfois en de multiples copies en tandem ou dispersées dans les génomes, seront décrites dans le chapitre 2. Elles contribuent pour une part importante au niveau de redondance génétique et de duplication de gènes et leur étude, quoique essentielle pour comprendre la dynamique des génomes complexes et l'héritabilité de certains traits, en est encore à ses balbutiements. Les transposons et rétrotransposons seront présentés dans le chapitre 3 et leur rôle dans la génération de nouveautés génétiques sera détaillé. Dans la plupart des espèces, les centromères sont présents à raison d'un par chromosome. Ces éléments répétés très particuliers sont indispensables à la ségrégation correcte des chromatides sœurs lors des divisions cellulaires. Ils seront étudiés dans le chapitre 4 et nous verrons que les organismes holocentriques échappent à la règle en exhibant plusieurs dizaines de centromères par chromosome. À l'opposé des centromères, les télomères sont des séquences hautement répétées trouvées aux extrémités des chromosomes afin d'empêcher la perte d'information génétique. Leur séquence et leur structure varient énormément d'un organisme à l'autre et certaines espèces ont développé des télomères très originaux, faits d'éléments répétés en tandem. Ces notions seront étudiées dans le chapitre 5. Les G-quadruplex, ces structures secondaires de l'ADN provoquées par la répétition régulière de paires de bases GC, sont présents dans tous les génomes eucaryotes. Leur distribution ainsi que leur rôle dans la transcription et dans la

réplication de l'ADN seront abordés dans le chapitre 6. Les différents types d'ADN satellite, que l'on retrouve en grand nombre chez les eucaryotes et dont la fonction précise n'est pas toujours claire, seront décrits dans le chapitre 7. Nous verrons que bien que les génomes procaryotes en contiennent peu, certaines bactéries les utilisent comme camouflage pour échapper au système immunitaire de leur hôte. En restant dans le monde procaryote, nous finirons au chapitre 8 par l'étude fascinante d'un autre mécanisme de défense bactérien, dirigé contre ces autres ennemis que sont les plasmides et les bactériophages : le système CRISPR-Cas. L'acquisition de petits morceaux d'ADN répétés en tandem, provenant d'envahisseurs étrangers à la cellule, procure aux eubactéries et aux archées une ligne de défense robuste. Et aux généticiens du XXIe siècle une infinité d'outils pour manipuler à façon leurs génomes préférés.

I.4. Bibliographie

Annaluru, N., Muller, H., Mitchell, L.A., Ramalingam, S., Stracquadanio, G., Richardson, S.M., Dymond, J.S., Kuang, Z., Scheifele, L.Z., Cooper, E.M. *et al.* (2014). Total synthesis of a functional designer eukaryotic chromosome. *Science*, 344(6179), 55–58.

Britten, R.J. and Kohne, D.E. (1968). Repeated sequences in DNA. *Science*, 161, 529–540.

Dupressoir, A., Vernochet, C., Bawa, O., Harper, F., Pierron, G., Opolon, P., Heidmann, T. (2009). Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proceedings of the National Academy of Sciences*, 106(29), 12127–12132.

Dymond, J.S., Richardson, S.M., Coombes, C.E., Babatz, T., Muller, H., Annaluru, N., Blake, W.J., Schwerzmann, J.W., Dai, J., Lindstrom, D.L. *et al.* (2011). Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature*, 477(7365), 471–476.

Fédry, J., Liu, Y., Péhau-Arnaudet, G., Pei, J., Li, W., Tortorici, M.A., Traincard, F., Meola, A., Bricogne, G., Grishin, N.V. *et al.* (2017). The ancient gamete fusogen HAP2 is a eukaryotic class II fusion protein. *Cell*, 168(5), 904–915.e10.

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

Jacob, F. (1977). Evolution and tinkering. *Science*, 196(4295), 1161–1166.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A. *et al.* (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate protokaryotype. *Nature*, 431(7011), 946–957.

Cette bibliographie est identique à celle de l'ouvrage correspondant en anglais publié par ISTE.

Muller, H. and Koszul, R. (2015). Conception et synthèse de néochromosomes. *Médecine thérapeutique/Médecine de la reproduction, gynécologie et endocrinologie*, 17(4), 228–236.

Ohno, S. (1972). So much "junk" DNA in our genome. *Evolution of Genetic Systems*, 23, 366–370.

Peterson, D.G., Schulze, S.R., Sciara, E.B., Lee, S.A., Bowers, J.E., Nagel, A., Jiang, N., Tibbitts, D.C., Wessler, S.R., Paterson, A.H. (2008). Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Research*, 12, 795–807.

Richard, G.-F., Kerrest, A., Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.*, 72(4), 686–727.

Thomas Jr., C.A. (1971). The genetic organization of chromosomes. *Annu. Rev. Genet.*, 5, 237–256.

Ullu, E. and Tschudi, C. (1984). Alu sequences are processed 7SL RNA genes. *Nature*, 312, 171–172.

Wolfe, K.H. and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387, 708–713.