

> Ces statistiques offrent un aperçu de l'évolution du paysage de la recherche avant et après le projet *Génome humain*. Cet aperçu montre l'importance accordée à un petit nombre de gènes codant des protéines «superstars», au détriment de travaux potentiellement intéressants sur d'autres gènes. Il y a eu aussi une bascule vers l'étude des régions non codantes du génome et vers la compréhension des interactions du matériel génétique et des protéines. Quant à la recherche pharmacologique, elle s'est concentrée sur un nombre limité de cibles protéiques.

Certaines de ces tendances sont familières aux biologistes, mais les quantifier et les visualiser permet de les considérer sous un nouveau jour.



22% des publications relatives aux gènes ne concernaient que 1% des gènes en 2017

Il n'y a pas d'histoire alternative où le projet *Génome humain* n'aurait pas existé pour comparer ce qu'il serait advenu sans. Il est donc impossible de dire si ces tendances en biologie seraient apparues de toute façon. D'autres facteurs, allant d'une puissance de calcul accrue à des méthodes de séquençage plus sophistiquées, ont également joué un rôle dans nombre de ces évolutions. Il est néanmoins clair que le catalogue du projet *Génome humain* a catalysé la révolution génétique en cours.

LES GÈNES SUPERSTARS

On pense souvent que le projet *Génome humain* a marqué le début de la recherche intensive de gènes codant des protéines. En fait, la publication de 2001 a signé la fin d'une quête qui a duré des décennies. En effet, la première preuve de l'existence d'un gène codant une protéine date de 1902, avec la découverte de l'hormone sécrétine cinquante ans avant que la structure de l'ADN ne soit découverte, et soixante-quinze ans avant que le séquençage du génome ne devienne courant [à cette époque, on ne formulait pas encore les choses en ces termes: les auteurs de l'étude, les physiologistes britanniques William Bayliss et Ernest Starling, ont rapporté que des cellules épithéliales de l'intestin produisaient une substance, la sécrétine, à partir d'un précurseur présent dans les

cellules et qui résistait à divers traitements chimiques; on a compris plus tard que la sécrétine est une protéine et son précurseur un gène, ndt]. Notre analyse montre que, entre le début du projet *Génome humain* en 1990 et son achèvement en 2003 (après la première publication en 2001), le nombre de gènes humains découverts (ou «annotés») a augmenté de façon spectaculaire. Il s'est soudain stabilisé au milieu des années 2000 à environ 20000 gènes codant des protéines, bien loin des 100000 estimés précédemment par de nombreux biologistes.

Bien que la découverte de gènes codant des protéines ait atteint un plateau, l'intérêt pour les gènes individuels a vite augmenté à la suite du projet *Génome humain*. Chaque année depuis 2001, entre 10000 et 20000 articles mentionnant des gènes codant des protéines ont été publiés.

Cependant, cet intérêt s'est principalement focalisé sur seulement quelques gènes. Avant 1990, le gène *HBA1* était le plus étudié, car il code l'une des protéines de l'hémoglobine. À partir de 1990, l'attention (d'après le nombre cumulé de publications) s'était ensuite portée sur le gène *CD4*, impliqué dans le fonctionnement des lymphocytes T et leur vulnérabilité au virus du sida. Mais l'intérêt pour ces deux gènes fait pâle figure à côté de l'explosion des articles sur des gènes individuels qui a suivi la publication du projet *Génome humain* en 2001. Certains gènes superstars, dont *TP53*, *TNF* et *EGFR*, ont fait l'objet de centaines d'articles par an, tandis que la plupart des autres gènes ne recevaient que peu d'attention. En 2017, 22% des publications relatives aux gènes ne faisaient référence qu'à 1% des gènes.

Une étude intensive est bien sûr justifiée pour les gènes qui ont une importance biologique majeure. Un bon exemple est le gène *TP53*, crucial pour la croissance et la mort cellulaire, et dont l'altération ou l'inactivation conduit à des cancers. On trouve des variations de ce gène dans plus de 50% des séquences génétiques tumorales. Il est mentionné dans 9232 publications entre 1976 et 2017.

On pourrait supposer que plus on en sait sur les mêmes quelques gènes, plus on serait enclin à explorer le reste du génome. Mais c'est le contraire qui s'est produit au cours des deux dernières décennies: on a accordé plus d'attention à une poignée d'élus. Bien que cela ait été signalé comme un problème potentiel lors du dixième anniversaire de la publication du projet *Génome humain*, il n'y a eu aucun changement de cap depuis.

Nos travaux antérieurs sur d'autres systèmes très différents, des réseaux humains à internet, indiquent que ce déséquilibre peut s'expliquer par une dynamique de type «plus on est riche, plus on s'enrichit». Il est probable qu'à mesure que le nombre de publications >

> focalisées sur le gène *TP53* augmente, il devient d'autant plus facile d'obtenir des financements, un encadrement, des outils et d'être cité pour les travaux futurs sur *TP53*. Dans la science des réseaux, ce phénomène est appelé « attachement préférentiel ». De fait, on constate que le nombre de nouvelles publications annuelles portant sur un gène donné est directement proportionnel à l'abondance de la littérature pré-existante sur ce gène.

Un défi pour la biologie consiste maintenant à déterminer ce qu'il faudra étudier ensuite. Les chercheurs consacrent-ils de l'argent, du temps et des efforts à ce qui est le plus important et le plus urgent, ou à ce qui est en terrain connu, parce que c'est une recette fiable pour obtenir des subventions et de la reconnaissance ?

DES TRÉSORS DANS LES POUBELLES

Un grand débat a précédé le lancement du projet *Génome humain* : était-il utile de cartographier les vastes régions non codantes du génome, que l'on appelait alors « ADN poubelle » ou « matière noire » du génome ? Grâce en grande partie au projet *Génome humain*, on sait maintenant que la majorité des séquences fonctionnelles du génome humain ne codent pas de protéines. Au contraire, des éléments tels que les ARN non codants, les promoteurs, les séquences stimulatrices [de petits segments d'ADN nécessaires pour activer les gènes, *ndt*] et d'innombrables motifs de régulation des gènes travaillent ensemble pour donner vie au génome. Les variations dans ces régions n'altèrent pas les protéines, mais peuvent perturber les réseaux qui régissent l'expression de ces dernières.

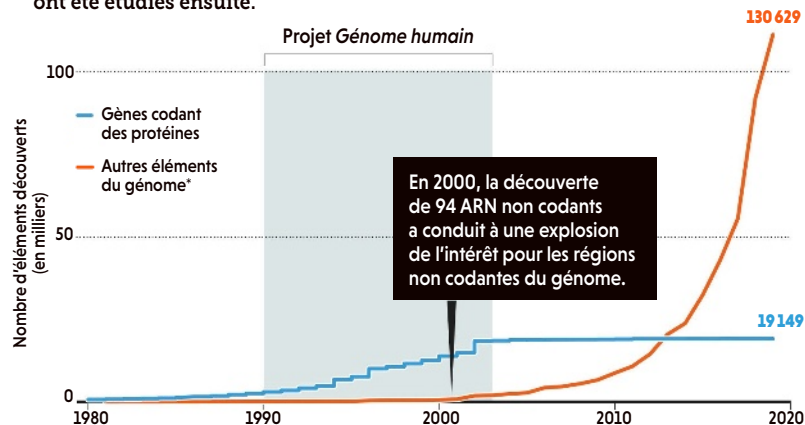
Avec la carte du génome humain en main, la découverte d'éléments non codants a explosé. Jusqu'à présent, le rythme a été cinq fois plus rapide que celui de la découverte des gènes codant des protéines et il ne montre aucun signe de ralentissement. De même, le nombre de publications portant sur ces constituants du génome a augmenté au cours de la période couverte par notre ensemble de données (1900 à 2017). Par exemple, il existe des milliers d'articles sur les ARN non codants, qui régulent l'expression des gènes.

Le projet *Génome humain* a également offert un moyen de cataloguer les variations génétiques humaines, y compris celles des polymorphismes touchant un seul nucléotide, les SNP. D'autres efforts importants ont permis de réduire le coût de l'identification de variations communes à des milliers d'individus, notamment celui du projet international *HapMap* (dont la troisième et dernière phase s'est achevée en 2010) et du projet *1000 génomes* (achevé en 2015). Ces ensembles de données, combinés aux progrès de l'analyse statistique, ont

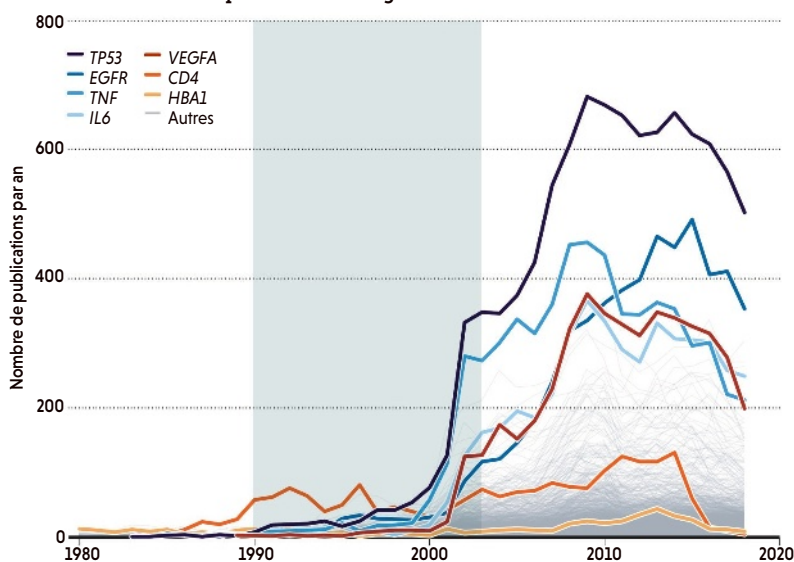
VINGT ANS DE POUBELLES, DE STARS ET DE MÉDICAMENTS

Ce que les chercheurs en génomique ont étudié, quand et pourquoi, tel que le retrace l'analyse bibliométrique.

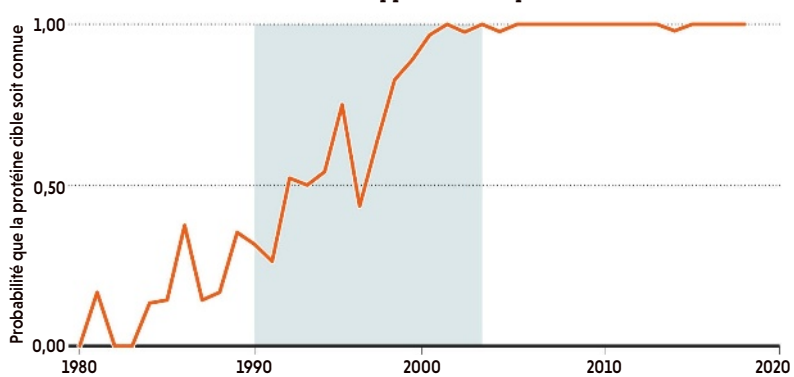
Éléments non codants. La plupart des gènes codant des protéines ont été découverts avant la première publication du projet *Génome humain* en 2001. De nombreux autres éléments du génome, auparavant nommés « ADN-poubelle », ont été étudiés ensuite.



Gènes stars. Le projet *Génome humain* a conduit à une explosion des recherches sur un petit nombre de gènes.



Cibles thérapeutiques. Depuis 2001, on a identifié les protéines cibles de près de 100 % des nouveaux médicaments approuvés chaque année aux États-Unis.



* Polymorphismes touchant un seul nucléotide (SNP), pseudogènes, ARN non codant, promoteurs, etc.

© Nature ; source : Barabási lab