

## L'ESSENTIEL

> En 2001, deux équipes ont indépendamment publié la première carte du génome humain, fournissant ainsi la liste de tous les gènes codant des protéines chez l'humain.

> Depuis, les recherches sur le génome humain ont explosé, notamment sur les régions du génome qui ne codent pas de protéines.

> L'analyse des publications scientifiques associées révèle qu'une grande partie concerne un petit nombre de gènes et qu'il reste encore de nombreuses cibles potentielles de médicaments à exploiter.

## LES AUTEURS

ALEXANDER J. GATES  
bio-informaticien,  
chercheur associé  
à l'université Northeastern,  
à Boston, aux États-Unis

DEISY MORSELLI GYSI  
bio-informaticienne en  
postdoctorat à l'université  
Northeastern, à Boston,  
aux États-Unis

MANOLIS KELLIS  
professeur d'informatique  
à l'institut de technologie du  
Massachusetts, à Cambridge,  
aux États-Unis

ALBERT-LÁSZLÓ BARABÁSI  
professeur de science  
des réseaux à l'université  
Northeastern, à Boston,  
aux États-Unis

Cet article est une traduction de « *A wealth of discovery built on the Human Genome Project – by the numbers* », publié sur *Nature.com* le 10 février 2021.

Le séquençage du génome humain a eu un impact considérable sur la génomique en termes de publications scientifiques, de nouveaux médicaments et de compréhension des maladies.

**L**e vingtième anniversaire de la publication de la première carte du génome humain (indépendamment par une équipe internationale autour du biologiste américain Craig Venter et par le Consortium international pour le séquençage du génome humain) offre l'occasion d'examiner en quoi ce projet a aidé la recherche des racines génétiques de maladies, favorisé la découverte de nouveaux médicaments et contribué à revoir notre compréhension du génome.

Pour illustrer cet impact, nous avons combiné plusieurs ensembles de données afin de quantifier les différents éléments génétiques qui ont été découverts et ont donné lieu à des publications, et visualiser comment le rythme et les sujets de publication ont évolué au fil des ans. Notre analyse rassemble des données telles que 38 546 transcriptions d'ARN, environ 1 million de polymorphismes touchant un seul nucléotide [les nucléotides sont les constituants de l'ADN, ndt] ou SNP, 1 660 maladies ayant des racines génétiques documentées, 7 712 produits pharmaceutiques autorisés ou expérimentaux et 704 515 articles scientifiques publiés entre 1900 et 2017.

Les résultats mettent en évidence comment le projet *Génome humain*, avec la liste exhaustive des gènes codant des protéines qu'il a permis d'établir, a marqué le début d'une nouvelle ère – celle de l'étude de la partie non codante du génome – et a ouvert la voie à des développements thérapeutiques. Surtout, ces résultats

permettent de suivre l'émergence d'une vision systémique de la biologie, parallèlement à la perspective classique focalisée sur chaque gène individuel, à mesure que les chercheurs cartographiaient les interactions des éléments constitutifs des cellules.

## UN APERÇU DES PUBLICATIONS DU DOMAINE

Notre analyse a cependant ses limites. Par exemple, il n'existe pas de consensus sur l'endroit où un gène commence et se termine ni même, étonnamment, sur la séquence exacte codant certains gènes. De multiples conventions sont utilisées en parallèle pour nommer certains éléments du génome, de sorte que notre méthodologie a parfois échoué à les relier. Et d'autres liens entre publications et éléments génomiques n'ont peut-être pas été ajoutés aux bases de données par leurs auteurs. Enfin, nos graphiques s'arrêtent en 2017, car il peut y avoir un décalage entre la publication d'un article et son entrée dans les bases de données que nous avons utilisées.

Toutefois, nous ne pensons pas que ces limitations modifient de façon notable les tendances globales constatées dans l'évolution de la recherche sur le génome au fil du temps. Et ces tendances persistent lorsque nous prenons en compte la croissance de la production scientifique en biologie au cours de la même période. Nous n'avons pas effectué ces contrôles sur toute la période qui a suivi la découverte des gènes, mais nous estimons que cela n'aurait pas modifié nos conclusions. >

**Lacunes**  
3% des gènes  
ne font l'objet  
d'aucune  
publication.

Le gène *ADRA1A* est la  
cible de 99 médicaments  
différents, dont 5% sont  
approuvés. Il est le sujet  
de 130 publications.

Le gène *TNF* est  
associé à 160 maladies  
connues, plus que tout  
autre gène.

## UN IMPACT PROFOND

Dans le schéma ci-contre, les 19 757 gènes qui codent des protéines sont disposés selon leur position relative le long de chacun des chromosomes, représentés par les anneaux. Le plan matérialise la publication des données du projet *Génome humain* en 2001. La profondeur des pics sous ce plan est proportionnelle au nombre de publications impliquant un gène donné depuis cette date ; la hauteur au-dessus du plan représente les publications antérieures. La largeur de la base de chaque pic reflète le nombre de maladies associées à chaque gène. Une poignée de gènes, distribués parmi tous les chromosomes, ont été abondamment étudiés. Il en est de même des régions non codantes situées entre les gènes codant des protéines (non représentées ici). Ces vingt dernières années, les chercheurs ont appris que ces régions non codantes jouent un rôle clé pour réguler la dynamique de l'expression des gènes.

### Une longue histoire

Le gène *TP53*, situé sur le chromosome 17, a été découvert en 1979. Associé à la plupart des cancers, il totalise 9 232 publications à ce jour.

### Les 8 gènes les plus étudiés

1. *TP53*
2. *TNF*
3. *EGFR*
4. *IL6*
5. *VEGFA*
6. *APOE*
7. *TGFB1*
8. *MTHFR*

### Nombre de maladies

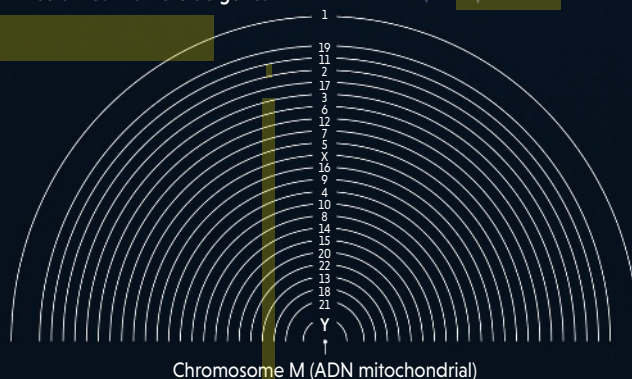
10 50 150

50 100 200  
100 500 1000

Nombre de  
publications  
avant 2001

Nombre de  
publications  
après 2001

Chromosomes ordonnés  
selon leur nombre de gènes



> 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 >