INSTITUT DE FRANCE
Académie des sciences

Comptes Rendus
Biologies

Thomas Bourgeron

Le projet gnomAD montre l'importance de ne pas avoir qu'un seul génome humain de référence !

Volume 343, issue 2 (2020), p. 123-125

Published online: 9 October 2020

https://doi.org/10.5802/crbiol.14

Les Comptes Rendus. Biologies *sont membres du*
Centre Mersenne pour l'édition scientifique ouverte
www.centre-mersenne.org
e-ISSN : 1768-3238

Comptes Rendus
Biologies
2020, 343, nº 2, p. 123-125
https://doi.org/10.5802/crbiol.14

C'est apparu dans la presse / *News and Views*

Le projet gnomAD montre l'importance de ne pas avoir qu'un seul génome humain de référence !

The gnomAD project and the importance of having more than only one reference human genome!

Thomas Bourgeron [a]

[a] Human Genetics and Cognitive Functions, Institut Pasteur, UMR3571 CNRS, Université de Paris, Paris, 75015, France

Courriel : Thomas.bourgeron@pasteur.fr

Manuscrit reçu et accepté le 29 juillet 2020.

On a l'habitude de se représenter le génome humain comme un long filament d'ADN homogène (3,2 milliards de nucléotides : A, T, G et C) et la mutation comme un élément perturbateur de cette parfaite organisation génétique. Les avancées récentes nous montrent au contraire que la mutation est plutôt la règle que l'exception. Chaque personne possède un génome garni d'environ 3 millions de variations génétiques plus ou moins grande, plus ou moins fréquentes et dont certaines sont considérées comme délétères pour le fonctionnement des gènes.

Le projet génome humain a été initié dans les années 80 et a permis d'établir une première séquence quasiment complète en 2001. Ce génome de référence est depuis mis à jour régulièrement afin de corriger des erreurs d'assemblage et de compléter les parties manquantes. Cependant, cette séquence de référence ne nous renseigne pas sur les différences génétiques entre les individus. Pour cela, plusieurs projets comme HapMap, 1000 génomes et ExAC ont identifié des différences génétiques individuelles permettant à la fois de retracer l'histoire des populations humaines et de mieux comprendre la vulnérabilité individuelle aux maladies génétiques.

Dans ce mouvement, le consortium Genome Aggregation Database (gnomAD) dirigé par Daniel G. MacArthur (Broad Institute) a récemment réalisé le plus grand catalogue de variations génétiques en agrégeant des données de 15708 génomes et de 125 748 exomes (la séquence de seulement la partie codante des gènes). Cette base est ouverte au public[1] (https://gnomad.broadinstitute.org/) et contient 241 millions de variations génétiques alors que l'étude précédente ExAC n'en contenait « seulement » 7.4 millions. Ce nouveau catalogue gnomAD fournit aussi plus de 300 000 variations structurales qui modifient la séquence de référence sur plus de 50 nucléotides (jusqu'à plusieurs millions).

[1] Cette base ne contient pas de données permettant l'identification des participants de l'étude mais uniquement des données agrégées qui indiquent pour chaque région du génome combien de personnes portent les variations rapportées.

Les retombées du projet gnomAD sont nombreuses et plusieurs articles utilisant cette ressource sont disponibles en accès libre sur le site de la revue Nature (https://www.nature.com/immersive/d42859-020-00002-x/index.html).

Ces études montrent en particulier qu'il existe dans la population générale des personnes porteuses de mutations qui entrainent une perte de la fonction de certains gènes (le catalogue gnomAD répertorie 443 769 de ces variations) [1]. Pour 1815 gènes, les deux copies sont inactivées, suggérant que certaines personnes puissent tolérer la perte complète de la fonction de ces gènes. Une autre étude montre qu'il existe chez 3,9% des personnes de très grands remaniements de l'ADN (de plus d'un million de nucléotide pour certains) et que 0,13% des personnes seraient porteuses d'un variant structural ayant tous les critères pour entrainer une maladie [2].

Cette nouvelle ressource ouvre des pistes pour identifier des traitements de maladies génétiques mais surtout elle permet dès maintenant de mieux interpréter les tests génétiques en comparant les résultats de séquençage obtenus chez les patients non plus sur un seul génome de référence mais sur des milliers et bientôt des millions de génomes. Cette étude oblige à revoir la vision simplificatrice d'un génome homogène porteur d'une mutation qui entraine une maladie. Le génome est variable et nous sommes dans certain cas capables d'accepter des mutations apparemment délétères. Pourquoi certaines personnes sont résilientes à la présence de ces mutations alors que d'autres vont déclarer une maladie reste encore largement incompris. D'autres projets regroupant des données génétiques et épidémiologiques, comme le projet UKBiobank [3], vont permettre de répondre à cette question mais ceci est une autre histoire...

English version

The human genome is usually thought of as a long filament of homogeneous DNA (3.2 billion nucleotides: A, T, G and C) and mutation as a disruptive element in this perfect genetic organization. Recent advances show us, on the contrary, that mutation is more the rule than the exception. Each person has a genome packed with about 3 million genetic variations, more or less large, more or less frequent and some of which are considered to strongly alter the gene function.

The Human Genome Project was initiated in the 1980s and established a first almost complete sequence in 2001. This reference genome has since been regularly updated to correct assembly errors and complete missing parts. However, this reference sequence does not tell us anything about the genetic differences between individuals. For this reason, several projects such as HapMap, 1000 Genomes and ExAC have identified individual genetic differences, shedding light on the history of human populations and on individual susceptibility to genetic diseases.

In this movement, the Genome Aggregation Database (gnomAD) consortium led by Daniel G. MacArthur (Broad Institute) recently completed the largest catalogue of genetic variations by aggregating data from 15,708 genomes and 125,748 exomes (sequence of only the coding part of genes). This database is open to the public[2] (https://gnomad.broadinstitute.org/) and contains 241 million genetic variations whereas the previous ExAC study contained "only" 7.4 million. This new gnomAD catalogue also provides more than 300 thousand structural variations that modify the reference sequence over 50 nucleotides (to millions).

The benefits of the gnomAD project are numerous and several articles using this resource are on open access at the Nature journal's website (https://www.nature.com/immersive/d42859-020-00002-x/index.html). These studies show that there are individuals in the general population who carry mutations that lead to a loss of function of certain genes (the gnomAD catalogue lists 443,769 of these variations) [1]. For 1,815 genes, both copies are inactivated, suggesting that some individuals may be able to tolerate the complete loss of function of these genes. Another study shows that 3.9% of people have very large DNA rearrangements (some with more than a million nucleotides) and

[2]There is no genetic data to identify study participants, only aggregate data that indicates for each region of the genome how many people had these variations.

that 0.13% of individuals carry a structural variant with all the criteria for causing disease [2].

This new resource opens up avenues for identifying treatments for genetic diseases but, above all, it already makes it possible to better interpret genetic tests by comparing the sequencing results obtained in patients no longer on a single reference genome but on thousands and soon millions of genomes. This study makes it necessary to review the simplifying vision of a homogeneous genome carrying a mutation that causes a disease. The genome is variable and we are in some cases able to accept apparently deleterious mutations. Why some people are resilient to the presence of these mutations and others will declare a disease is still largely unknown. Other projects gathering genetic and epidemiological data such as the UKBiobank project [3] will help answer this question, but that is another story...

Références / *References*

[1] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur, « The mutational constraint spectrum quantified from variation in 141,456 humans », *Nature* **581** (2020), nº 7809, p. 434-443.

[2] R. L. Collins, H. Brand, K. J. Karczewski, X. Zhao, J. Alföldi, L. C. Francioli, A. V. Khera, C. Lowther, L. D. Gauthier, H. Wang, N. A. Watts, M. Solomonson, A. O'Donnell-Luria, A. Baumann, R. Munshi, M. Walker, C. W. Whelan, Y. Huang, T. Brookings, T. Sharpe, M. R. Stone, E. Valkanas, J. Fu, G. Tiao, K. M. Laricchia, V. Ruano-Rubio, C. Stevens, N. Gupta, C. Cusick, L. Margolin, Genome Aggregation Database Production Team; Genome Aggregation Database Consortium, K. D. Taylor, H. J. Lin, S. S. Rich, W. S. Post, Y. I. Chen, J. I. Rotter, C. Nusbaum, A. Philippakis, E. Lander, S. Gabriel, B. M. Neale, S. Kathiresan, M. J. Daly, E. Banks, D. G. MacArthur, M. E. Talkowski, « A structural variation reference for medical and population genetics », *Nature* **581** (2020), nº 7809, p. 444-451.

[3] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, « The UK Biobank resource with deep phenotyping and genomic data », *Nature* **562** (2018), nº 7726, p. 203-209.