



**HAL**  
open science

## Le next-generation biobanking

Georges Dagher, Maria Luisa Lavitrano, Paul Hofman

► **To cite this version:**

Georges Dagher, Maria Luisa Lavitrano, Paul Hofman. Le next-generation biobanking: Un défi des données numériques. *Médecine/Sciences*, 2018, 34 (10), pp.849-851. 10.1051/medsci/2018203. hal-02970082

**HAL Id: hal-02970082**

**<https://hal.science/hal-02970082>**

Submitted on 17 Oct 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

► La recherche biomédicale connaît depuis le début du siècle un bouleversement de grande ampleur avec l'avènement de technologies à grand débit (les -omiques) appliquées à la biologie et associées à des approches biologiques, moléculaires ou aux techniques d'imagerie. Cette révolution méthodologique s'appuie sur l'analyse d'échantillons biologiques prélevés sur les patients puis conservés dans des biobanques. L'intégration des données massives obtenues par ces différentes technologies et leur analyse devrait permettre d'accroître nos connaissances des mécanismes complexes des pathologies humaines et une meilleure stratification des patients selon une nomenclature génétique ou moléculaire. L'accroissement exponentiel des données générées et leur complexité nécessitent cependant la mise en place d'infrastructures adaptées, de nouvelles modalités d'accès et d'échanges de ces données ainsi qu'une organisation optimisée des biobanques afin d'intégrer de nouvelles disciplines adaptées à l'analyse de ces données. ◀

Le nombre de sujets atteints de pathologies à dimension globale ou systémiques ou de pathologies chroniques a nettement augmenté au cours des dernières décennies. Ceci est illustré par le Sida (syndrome d'immunodéficience acquise), le SRAS (syndrome respiratoire aigu sévère), la résurgence d'infections comme la dengue, la fièvre jaune, le chikungunya, ou encore par l'accroissement constant du nombre de patients atteints de maladies métaboliques ou de cancer. L'impact de ces pathologies en termes de santé publique et de conséquences sociales et économiques est en fait dramatique. Le vieillissement de la population mondiale fait également de ces pathologies un véritable fléau. Pour répondre aux enjeux biomédicaux et socio-économiques que ces maladies posent et mieux les combattre, la recherche déploie de nouvelles technologies fondées

## Numérique et santé (9) Le next-generation biobanking

### Un défi des données numériques

Georges Dagher<sup>1</sup>, Maria Luisa Lavitrano<sup>2</sup>,  
Paul Hofman<sup>3</sup>



<sup>1</sup>Biobanques, Inserm US013, Hôpital de la Salpêtrière, 47, boulevard de l'Hôpital, 75651 Paris, France.

<sup>2</sup>Université Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italie.

<sup>3</sup>Université Côte d'Azur, BB-0033-00025, Campus Valrose, 28, avenue de Valrose, 06108 Nice, France.

[georges.dagher@inserm.fr](mailto:georges.dagher@inserm.fr)

sur des analyses d'échantillons biologiques provenant de patients. Ces échantillons sont collectés, transformés et conservés dans des structures dédiées : les biobanques. En utilisant ces ressources, la génomique a permis des progrès significatifs au cours des dernières décennies [1, 2] concourant à une meilleure compréhension de la pathogénie et de l'épidémiologie d'agents microbiens, notamment dans le cas de la méningite, de la maladie du légionnaire, de la diphtérie, de la fièvre typhoïde, de la grippe, du paludisme, de la dengue, du SIDA, etc. Une des clés principales de ce progrès a été le développement de la génomique, qui a permis de préciser les variations génétiques des agents pathogènes et de révolutionner ainsi le développement de vaccins.

Le séquençage du génome humain, réalisé en 2001, a ouvert la voie à l'identification de gènes spécifiques impliqués dans des pathologies particulières et a constitué le fondement d'une stratégie qui consiste à rechercher et analyser des variants génétiques qui prédisposent l'individu au développement de ces pathologies. Les études *genome wide scan* ont ainsi permis d'identifier des associations génétiques à des pathologies complexes. La clé de ce succès a été la mise en place de *consortiums* internationaux et de méta-analyses portant sur de grandes séries de patients permettant ainsi de comprendre la complexité des maladies et de détecter les variants génétiques « d'effet modeste ». Cette stratégie s'est avérée fructueuse pour de nombreuses maladies chroniques, comme le diabète de type 1 [3] et de type 2 [4, 5], les coronaropathies [6, 7], les cancers du sein [8, 9], du côlon [10, 11] et de la prostate [11, 12], la dégénérescence maculaire liée à l'âge [13, 14], la maladie de Crohn [7, 15], l'autisme [16], les maladies neurodégénératives.

Cette stratégie de recherche a abouti à la promesse d'une « médecine personnalisée » ou « stratifiée » qui repose notamment sur la prévision du risque de développer une pathologie, et ce sur la base de données génomiques, le choix du traitement approprié, l'évaluation de la réponse et des effets secondaires liés à la thérapie. Certaines instances et certains chercheurs considèrent ainsi la médecine personnalisée comme une médecine génomique.

Dans son dernier rapport, l'*European science foundation* (ESF) note cependant que l'étendue de l'information qui pourrait être utile dans la définition et le diagnostic d'une pathologie est bien plus vaste que celui des données génomiques. La définition par cette instance de la médecine personnalisée inclut les différences individuelles à tous les stades de la pathologie, de la prévention au traitement et à son suivi. Les différences génétiques peuvent en effet déterminer la prédisposition des individus à certaines pathologies et leur réponse thérapeutique. Cependant, la plupart des pathologies sont polygéniques. Elles peuvent aussi être influencées par une diversité de facteurs dits « environnementaux ». La pathologie et le tableau clinique sont ainsi la résultante de l'interaction entre gènes et environnement. Les différences de caractéristiques biologiques, biochimiques, physiologiques sont donc à prendre en considération dans le choix du traitement personnalisé.

L'Établissement français du sang (EFS) préconise ainsi le recueil de plusieurs types de données : environnementales, génomiques, épigénomiques, protéomiques, métabolomiques, lipidomiques, microbiome et imagerie. Ces données nécessitent d'être intégrées, analysées et interprétées afin de mieux préciser le diagnostic et de conduire à un traitement approprié au patient. Le terme « médecine de précision » a été ainsi introduit pour préciser et élargir ceux déjà largement utilisés de « médecine personnalisée » ou de « médecine guidée par la génomique ». La médecine de précision implique une meilleure compréhension de l'étiologie de la maladie et des mécanismes moléculaires afin d'asseoir le bien-fondé de la future thérapeutique. Garraway *et al.* [17] en propose une définition fondée sur trois critères : (1) une résolution moléculaire accrue, (2) une clarté mécanistique et (3) une thérapeutique bien fondée.

Au cours de la dernière décennie, la médecine de précision a investi progressivement différentes pathologies communes complexes, parmi lesquelles l'oncologie en constitue l'avant-garde [18]. On retrouve ainsi aujourd'hui dans différents domaines de recherche médicale des jeux de données associées à des ressources biologiques pour lesquelles les observations sont décrites selon différentes modalités (-omiques, biochimie, environnementales, etc.). Les observations histopathologiques et radiologiques, analysées et interprétées visuellement par des experts, sont désormais de plus en plus quantifiées [19, 20]. Le défi actuel réside dans l'analyse de ces données très complexes et l'extraction de nouvelles connaissances. La progression d'une pathologie ou d'une tumeur est un processus biologique dynamique et évolutif. Il implique différents tissus et biofluides avec un génome façonné par des aberrations de gènes, par des facteurs épigénétiques, un contexte biologique cellulaire, et des données spécifiques au mode de vie de l'individu et à son environnement [21, 22]. Des méthodes statistiques

et mathématiques sophistiquées ont été développées pour l'analyse, l'interprétation et la validation des données biologiques. Des techniques et outils de calcul sont ainsi en constant développement. Cette approche a permis en particulier d'identifier les vulnérabilités et les processus mis en œuvre dans le développement des tumeurs [21, 23].

La multiplicité des spécialités nécessaires pour générer ces données, leur intégration et leur analyse, nécessitent de repenser l'organisation et la structuration des biobanques afin d'y rattacher les spécialistes de différents domaines : médecine, biologie, imagerie, statistique, bioinformatique, mathématiques, etc. Cette évolution nécessiterait, au niveau national et international, une liaison étroite avec les biobanques déjà impliquées dans la maîtrise de la collecte, l'intégration et l'analyse des données, ce qui participerait à la mise en place d'une nouvelle génération de biobanques [24, 25]. Elle devrait associer les bases de données capables d'héberger des données biologiques, de -omiques, phénotypiques, d'imagerie, de *radiomiques*, mais aussi comportementales et familiales provenant de différents centres à des fins d'analyses [26]. Elle nécessite de relever plusieurs défis. Le premier est celui d'infrastructures de calcul capables de traiter des téra, penta et exabytes [27]. À ceci s'ajoute le besoin de développement d'une suite de méthodes d'analyses et d'algorithmes appropriés. Cette évolution s'accompagnera également d'un changement de paradigme de l'analyse de données provenant d'un grand nombre de patients à celle d'un grand volume de données qui, lui, sera issu d'un seul sujet. Ainsi la notion de *fat data* pourra progressivement compléter les approches portant sur les données massives ou de *big data*. Il devient ainsi important de développer les formations et les compétences sur le *deep learning* et de créer de nouveaux champs d'expertise pluridisciplinaires. ♦

## SUMMARY

### Next generation biobanking: the challenge of data

Human biological samples are key resources in unravelling physiopathological factors underlying diseases and influencing their outcome. By making use of these resources, genomics, proteomics and molecular imaging techniques have achieved unprecedented progress in the past decade. The development of genomics platforms, molecular imaging as well as bioinformatics allowed a significant development of the biomarkers field thus realizing significant advances towards personalized medicine. The exponential increase of data, their complexity, the necessity of their integration for analysis require the development of appropriate infrastructures.

