Les *Big data* en oncologie : de la recherche fondamentale à des applications au bénéfice du patient

Par Emmanuel BARILLOT
Directeur de l'unité INSERM 900 à l'Institut Curie
et Philippe HUPÉ
Directeur adjoint de la plateforme de bioinformatique de l'Institut Curie

L'oncologie est aujourd'hui intimement liée au numérique, et ce, aussi bien pour les activités de recherche que pour les soins. Ce mariage est issu avant tout de notre capacité à explorer la dimension moléculaire des cellules grâce au séquençage de leur génome. Les mutations mises en évidence sont dès lors autant de cibles thérapeutiques potentielles pour ces nouvelles molécules pharmaceutiques nommées inhibiteurs ciblés qui offrent des perspectives prometteuses de personnalisation des traitements. Aussi assiste-t-on à la mise en place de plans nationaux de médecine de précision prévoyant le séquençage de centaines de milliers, voire de millions de génomes humains dans les années à venir (et donc la génération de pétaoctets de données), et à l'émergence dans les pays en avance dans le domaine d'une filière industrielle qui intéresse les grands acteurs du numérique et ses *start-ups*.

Big data et cancer : les enjeux

Le cancer est depuis 2007 la première cause de décès en France métropolitaine (30 %) comme dans beaucoup d'autres pays développés économiquement, juste devant les maladies cardiovasculaires (28,9 %), et le nombre des cas est en forte augmentation dans beaucoup de pays.

De fait, lutter contre cette maladie complexe est un enjeu majeur de santé publique. Parmi les urgences de la recherche en oncologie, nous mentionnerons l'épidémiologie qui vise à identifier les facteurs de risque de cancer, l'identification de nouveaux facteurs pronostiques et prédictifs pour le traitement, la découverte de nouvelles cibles thérapeutiques et la conception de nouveaux médicaments et de nouvelles stratégies thérapeutiques.

L'avènement des technologies dites à haut débit a permis depuis plusieurs années de mieux comprendre les mécanismes moléculaires impliqués dans la progression tumorale. Aujourd'hui, c'est un véritable arsenal de technologies de pointe qui s'offre aux chercheurs et aux médecins pour caractériser de plus en plus finement les tumeurs au niveau moléculaire (en étudiant le génome avec l'ADN, l'ARN, le protéome avec les protéines et leurs modifications, ou l'épigénome [1]) ou au niveau morphologique, grâce à des techniques d'imagerie biomédicale de plus en plus sophistiquées.

L'apport de ces outils est double : cognitif, puisqu'ils permettent de mieux comprendre les mécanismes de la progression tumorale, et pratique, puisqu'ils ouvrent la voie à l'amélioration de la prise en charge thérapeutique.

Partant du principe que chaque tumeur cancéreuse est unique, l'utilisation de ces différentes technologies pour en caractériser la carte d'identité moléculaire permet de tailler sur mesure le traitement pour chaque patient, individuellement, au vu des anomalies observées, et de sélectionner la thérapie la plus efficace pour éradiquer la maladie. C'est ce que l'on appelle la médecine de précision.

Big data et cancer : les données

Le cancer est une maladie du génome : il est la conséquence d'une accumulation successive de mutations de l'ADN qui perturbent le paysage moléculaire et le fonctionnement de la cellule. En effet, la mutation d'un gène peut conduire

(1) L'épigénome se définit comme l'ensemble des modifications du génome autres que les mutations (substitution d'une base nucléique par une autre, réarrangements chromosomiques ou altérations du nombre de copies d'ADN) : il recouvre, par exemple, les méthylations de l'ADN (qui inactivent celui-ci), les modifications chimiques de ses protéines compagnonnes (la méthylation ou l'acétylation des histones, par exemple) ou encore la conformation tridimensionnelle de la chromatine (génome et protéines compagnonnes).

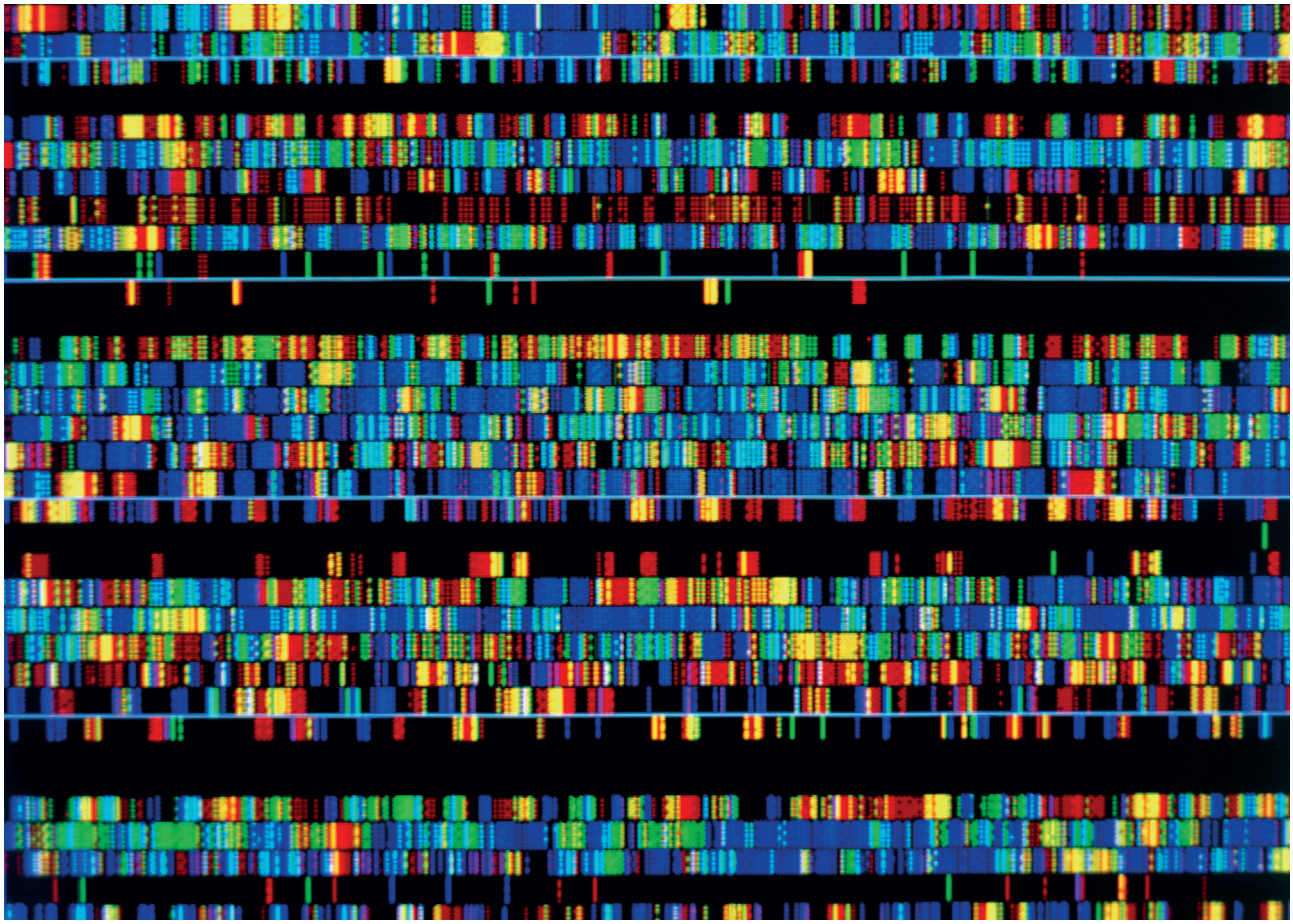

Photo © James King-Holmes/SPL-PHANIE

Une séquence d'ADN humain restituée sur un écran d'ordinateur sous la forme d'une série de segments de couleur.

« Le séquençage du génome humain a permis le développement d'un ensemble d'approches à haut débit rendant possible la caractérisation génétique des tumeurs. »

à la synthèse d'une forme anormale de la protéine correspondante qui ne sera pas fonctionnelle ou qui, au contraire, sera toujours active constitutionnellement.

Si plusieurs protéines sont affectées, des fonctions essentielles comme le contrôle de la prolifération cellulaire ou la mort cellulaire programmée seront perturbées, ce qui pourra conduire au développement d'une tumeur et à sa progression, éventuellement jusqu'au stade de la métastase. Depuis une vingtaine d'années, la pharmacopée s'est enrichie de molécules thérapeutiques ciblant spécifiquement certaines protéines fréquemment mutées dans les cancers, les inhibiteurs ciblés, qui font l'objet de nombreux programmes de développement et dont le nombre est en constante augmentation. Cela explique l'importance de l'utilisation des technologies basées sur le séquençage d'ADN dans la lutte contre le cancer, aussi bien en matière de recherche fondamentale qu'en matière de recherche clinique.

Le séquençage du génome humain (programme *Human Genome Project* initié en 1990 et achevé en 2003) a constitué une étape décisive en fournissant une description complète d'un génome humain, et en ouvrant ainsi la voie vers de nouveaux champs d'exploration. La bioinformatique y a joué un rôle crucial pour aider à l'assemblage des trois milliards de paires de bases d'ADN qui constituent ce génome. Le séquençage du génome humain a permis le développement d'un ensemble d'approches à haut débit rendant possible la caractérisation génétique des tumeurs. Parmi celles-ci, nous citerons la technologie des biopuces ou "*microarrays*", qui sont apparues en 1995 et dont l'utilisation est désormais supplantée par le séquençage à haut débit apparu en 2005. Cette technologie permet de séquencer de manière massive et en parallèle des molécules d'ADN présentes dans un échantillon, et ce en un temps et pour un coût réduits. Alors que le séquençage du premier génome humain avait nécessité 13 années de travail pour un coût de 3 milliards de dollars, il est désormais possible de séquencer un génome complet en 3 jours seulement, pour environ 1 000 dollars. La société Illumina domine ce marché sur lequel d'autres acteurs (comme Oxford Nanopore Technologies ou Pacific Biosciences) tentent de se positionner. La machine HiSeq X Ten d'Illumina permet de séquencer 18 000 génomes humains par an, moyennant toutefois un investissement de 10 millions de dollars pour l'acquisition des séquenceurs. L'apparition de ces technologies marque véritablement l'entrée de la biologie et de la cancérologie dans l'ère des

données massives, ou *Big data*. En effet, cette machine génère une quantité de données quotidienne de plusieurs dizaines de téraoctets. Une fois encore, non seulement la bioinformatique, mais aussi l'informatique et les technologies de l'information sont autant de maillons clés pour relever les défis posés par l'arrivée de ces données dont la volumétrie et la complexité défient notre capacité à les traiter et les comprendre.

Big data et cancer : les défis

Réussir à analyser ces données biomédicales présuppose de relever plusieurs défis majeurs.

Le premier est d'ordre technique. En effet, il est nécessaire de construire l'architecture informatique capable de centraliser dans un système d'information sécurisé et unifié toute cette myriade de données biomédicales, en respectant scrupuleusement leur diversité, leur hétérogénéité et leur complexité : cela va des rapports de consultation ou d'examen clinique (généralement en volumes restreints, mais non codifiés et exprimés en langage naturel) aux données génomiques (bien décrites, mais très volumineuses). La solution doit bien évidemment garantir le passage à l'échelle pour pouvoir absorber une quantité toujours croissante de données. Le recours à des matériels informatiques assurant le stockage et le calcul à haute performance est donc indispensable et l'on voit se multiplier des technologies, comme Hadoop, MPI ou NoSQL, qui permettent la gestion et le traitement de gros volume de données non structurées.

Le second défi est scientifique : analyser ces données suppose le développement et l'application de méthodologies mathématiques et statistiques sophistiquées qui soient capables de combiner dans un seul et même modèle l'ensemble des informations afférentes à chaque patient, que ces données soient cliniques, biologiques, moléculaires ou d'imagerie.

Le troisième défi est organisationnel : la production des données afférentes à un même patient fait intervenir de multiples acteurs, chacun d'eux produisant des données en des temps et en des lieux potentiellement éloignés, et avec des objectifs, des pratiques, des contraintes et des habitudes qui sont propres à chacun d'eux (médecins oncologues, infirmières, personnel des plateformes d'imagerie et d'analyses biologiques, biologistes, bioinformaticiens, informaticiens médicaux...). Par conséquent, la standardisation et l'harmonisation des procédures entre les différents acteurs constituent un passage obligé pour en garantir la qualité. Mais c'est aussi un passage complexe, car il implique un changement majeur dans leurs habitudes de travail. À l'évidence, l'intégration des données biomédicales est une aventure collective et multidisciplinaire nécessitant un important travail d'équipe dans lequel chacun des acteurs doit comprendre les bénéfices qu'il pourra en tirer en termes d'efficacité.

Enfin, le dernier défi est juridique et éthique : garantir que les solutions mises en œuvre respectent et protègent la vie privée et l'intimité de chacun est essentiel. Comment préserver l'intimité et l'anonymat des patients, dans ces bases de données, sachant que la séquence d'ADN est par essence identifiante de la personne, et que quatre données spatiotemporelles suffisent à identifier une personne dans une base de données (comme l'ont montré les études d'Yves-Alexandre de Montjoye) ? Comment offrir cette garantie tout en tirant parti de la richesse d'information des données personnelles pour améliorer les soins, soit au sein de programmes de recherche, soit pour la prise de décision thérapeutique, par comparaison directe des données d'un nouveau patient à la base de référence des patients qui l'ont précédé, et dont le profil de réponse thérapeutique au traitement administré est connu ? En effet, des patients de plus en plus nombreux, conscients que leurs données médicales peuvent servir à d'autres et que leur propre traitement a lui aussi bénéficié de l'éclairage apporté par des cas antérieurs, souhaitent partager largement leurs données.

Big data et cancer : les projets

Toutes ces questions sont d'ores et déjà posées dans le cadre des différentes initiatives nationales et internationales visant à caractériser le génome de patients atteints d'un cancer. *The Cancer Genome Atlas* (2005) et *l'International Cancer Genome Consortium* (2008) ont été les premières initiatives de recherche systématique à grande échelle pour cataloguer les différents types de cancer, et ce grâce à des études rétrospectives.

En 2013, le Royaume-Uni a créé la société *Genomics England*, qui est entièrement contrôlée par le *Department of Health* et qui est chargée de réaliser le séquençage de 100 000 génomes de patients atteints de maladies génétiques ou de cancers pour orienter la prise de décision médicale. Depuis lors, les projets se sont multipliés dans le monde.

Ainsi, en janvier 2015, le président américain Barack Obama a lancé la PMI (*Precision Medicine Initiative*), dont le périmètre dépasse l'oncologie. Ce projet, qui prévoit de constituer une cohorte d'un million d'Américains, sera financé à hauteur de 215 millions de dollars en 2016.

Pour rivaliser avec ce plan, la Chine a annoncé en mars 2016 un programme similaire s'étendant sur 15 ans et doté de plusieurs milliards de dollars pour séquencer le génome de 100 millions de Chinois à l'horizon 2030.

Tous ces programmes ont en commun de viser en parallèle la construction d'une filière industrielle nationale de génomique et d'ouvrir à la recherche la formidable masse des données collectées, selon des conditions très strictes, notamment de confidentialité.

Après l'Angleterre, les États-Unis et la Chine, mais aussi le Qatar, le Danemark, les Pays-Bas, l'Australie, l'Estonie et bien d'autres encore, la France a annoncé, en juin 2016, le lancement de son programme « France Médecine Génomique 2025 » qui vise à séquencer l'équivalent de 60 000 génomes de porteurs de maladies génétiques rares et 175 000 génomes de tumeurs de personnes atteintes d'un cancer – par an – d'ici à 2020, couvrant ainsi tous les cancers métastatiques. Le modèle français de

Big data génomique est en accord avec les principes fondateurs de notre médecine nationale, puisqu'il prévoit un accès pour tous à cette technologie, à la différence des modèles économiques américains et anglais. Sa prise en charge par l'Assurance maladie est prévue, même si les conditions de celle-ci doivent encore être précisées. Enfin, l'organisation de la formation des différents acteurs de ces programmes, à commencer par celle des médecins prescripteurs, constitue l'un des maillons indispensables de ce plan.

Big data et cancer : les acteurs industriels

Le véritable tsunami de données auquel doit faire face la cancérologie oblige cette discipline à recourir à des technologies numériques de pointe pour collecter, traiter, résumer et extraire les informations pertinentes pour aider les cliniciens dans leurs décisions thérapeutiques. Des outils s'appuyant sur des techniques d'apprentissage statistique et de traitement automatique du langage naturel sont désormais utilisés pour produire de la connaissance à partir de l'ensemble des données biomédicales disponibles dans les dossiers de suivi de patients.

Mentionnons, par exemple, l'utilisation d'IBM Watson Health™ par l'Institut new-yorkais *Memorial Sloan-Kettering Cancer Center* ou celle du système CancerLinQ™ soutenu par l'ASCO (*American Society of Clinical Oncology*). À l'instar d'IBM, tous les gros acteurs du numérique se positionnent actuellement sur le marché de la santé et de la médecine de précision, comme Intel (et son "*Extreme-Scale Computing*" pour la médecine de précision), Google (avec son projet « Google Genomics »), ou Apple. Ce dernier propose son offre « ResearchKit », qui vise un autre aspect de la médecine de précision : le développement de la e-Santé, qui tire parti de l'apport des objets connectés dans le lien entre patients et médecins.

Citons, à ce titre, l'application mobile MoovCare™ développée par le Dr. Fabrice Denis pour assurer la télésurveillance de patients souffrant d'un cancer du poumon : la saisie hebdomadaire de symptômes, tels que toux, fatigue, dyspnée, etc. par le patient *via* son application, couplée à un algorithme de décision, permet d'alerter le médecin, qui peut, au besoin, appeler son patient et programmer une visite. Une étude clinique a permis de démontrer la réelle plus-value apportée par cette application en termes d'amélioration de la survie des patients. Tout cela démontre également qu'à l'ère du numérique et du patient digital, dans un monde où la masse des données personnelles croit de manière exponentielle, l'apport des technologies du numérique en e-Santé doit non pas conduire à une virtualisation de la relation médecin-patient, mais bien, au contraire, favoriser le lien humain et personnel qui existe entre les deux acteurs. Éviter l'écueil de la dépersonnalisation est assurément l'un des enjeux centraux de la médecine de précision, à l'ère du numérique.

Enfin, les *start-ups* fleurissent également sur le marché de la médecine de précision, que ce soit pour l'analyse des données cliniques en langage naturel (*Sword services* [2]), l'analyse génomique (*Integragen* [3]) ou les plateformes d'analyse bioinformatique pour la clinique (*Sophia Genetix* [4]).

Big data et cancer : les perspectives

Le boom du numérique dans le domaine de la santé ne fait que commencer et l'on peut parier sur une croissance ininterrompue, et pour longtemps. Les perspectives en recherche, en santé publique et en soins sont infinies. Les objets connectés, qu'ils soient liés à un individu ou qu'ils échantillonnent des données environnementales, vont rendre possibles des études épidémiologiques d'une variété inépuisable permettant de détecter des facteurs de risque. La caractérisation moléculaire systématique des patients continuera à orienter de plus en plus la recherche en oncologie et ses progrès rendront indispensable une caractérisation toujours plus fine en support à la prévention et aux soins. On sait, par exemple, dans le cas du cancer, que la tumeur est hétérogène – il faudra donc séquencer génétiquement plusieurs prélèvements tumoraux et non pas un seul, voire un grand nombre de cellules uniques –, que les mutations de son génome ne suffisent pas à prédire son devenir – il faudra donc aussi déterminer son profil d'expression génique, son épigénome, là encore par séquençage, et son protéome –, qu'elle interagit avec un microenvironnement (en particulier immunologique) qui est crucial pour son devenir – il faudra donc caractériser aussi les cellules voisines –, qu'elle évolue dans le temps, *a fortiori* quand elle est soumise à un traitement – il faudra donc répéter l'analyse à intervalles réguliers.

Ces caractérisations multi-échelles de la tumeur vont faire croître le volume des informations recueillies de plusieurs ordres de grandeur. Les technologies de mesure permettant de traiter ces données sont d'ores et déjà en grande partie disponibles, le parcours de soins est en train d'être adapté pour permettre ce type d'investigation [5], et les technologies numériques se mettent en place.

Les défis à relever à l'avenir portent avant tout sur notre capacité à comprendre ces données multi-échelles, à les intégrer dans un modèle mathématique cohérent de *patient virtuel* en vue de leur utilisation pour aider à la décision thérapeutique, ce qui reste encore du domaine de la recherche.

(2) http://www.sword-services.com
(3) http://www.integragen.com
(4) http://www.sophiagenetics.com
(5) Des essais cliniques de médecine de précision, comme l'essai SHIVA coordonné par le Dr. Le Tourneau à l'Institut Curie, ont permis de valider depuis plusieurs années la faisabilité de l'utilisation de la médecine de précision dans le traitement des patients cancéreux.