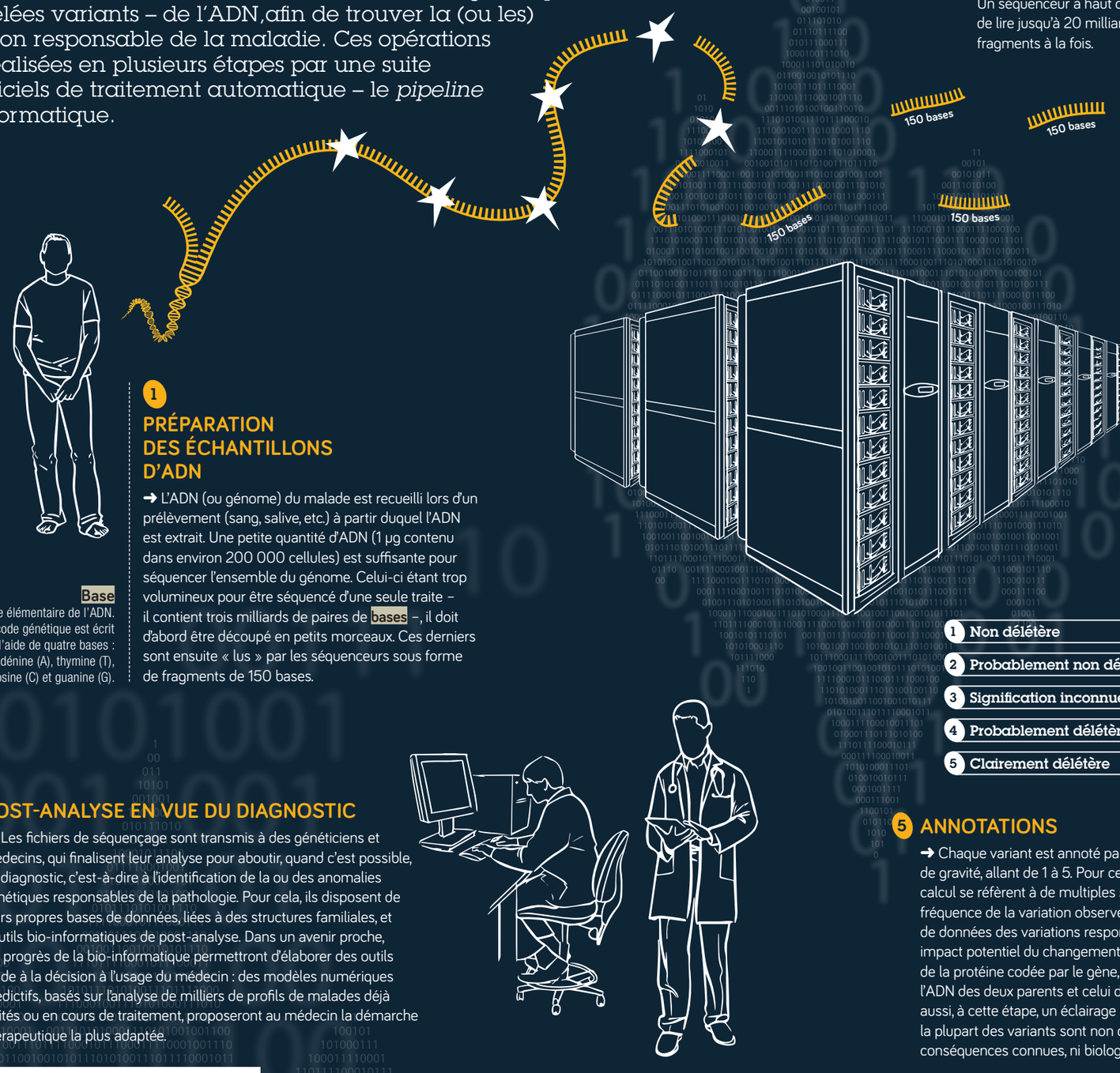
Pipeline bio-informatique

PRINCIPE

Diagnostiquer une maladie d'origine génétique passe par le séquençage à haut débit et l'analyse de l'ADN du patient. Cette opération permet de déchiffrer l'ordre (la séquence) des bases de l'ADN.
Il s'agit de détecter et classer toutes les anomalies génétiques – appelées variants – de l'ADN, afin de trouver la (ou les) mutation responsable de la maladie. Ces opérations sont réalisées en plusieurs étapes par une suite de logiciels de traitement automatique – le *pipeline* bio-informatique.



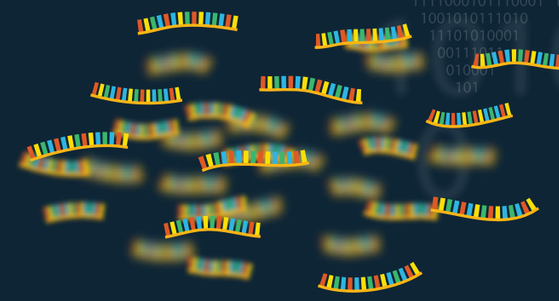
Base
Brique élémentaire de l'ADN. Le code génétique est écrit à l'aide de quatre bases : adénine (A), thymine (T), cytosine (C) et guanine (G).

1 PRÉPARATION DES ÉCHANTILLONS D'ADN

➜ L'ADN (ou génome) du malade est recueilli lors d'un prélèvement (sang, salive, etc.) à partir duquel l'ADN est extrait. Une petite quantité d'ADN (1 µg contenu dans environ 200 000 cellules) est suffisante pour séquencer l'ensemble du génome. Celui-ci étant trop volumineux pour être séquencé d'une seule traite – il contient trois milliards de paires de bases –, il doit d'abord être découpé en petits morceaux. Ces derniers sont ensuite « lus » par les séquenceurs sous forme de fragments de 150 bases.

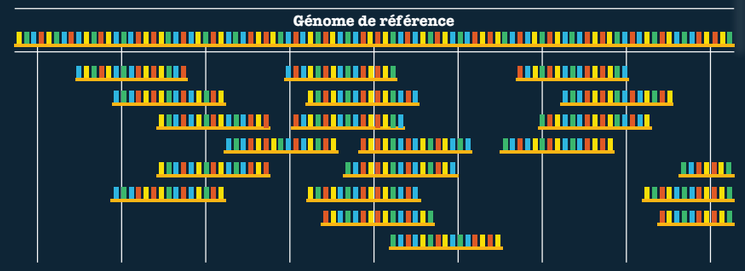
2 SÉQUENÇAGE À HAUT DÉBIT

➜ Des appareils à haut débit – les séquenceurs – déterminent l'ordre des bases de tous les morceaux d'ADN précédemment préparés : A (adénine), T (thymine), C (cytosine) et G (guanine). Un séquenceur à haut débit est capable de lire jusqu'à 20 milliards de ces fragments à la fois.



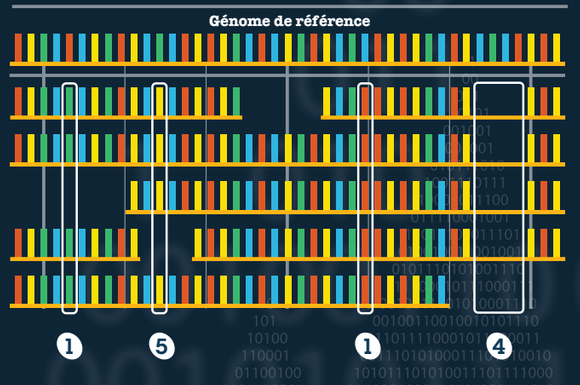
3 ALIGNEMENTS

➜ Les séquences obtenues doivent ensuite être remises dans le bon ordre par traitement bio-informatique, pour reconstituer le génome du patient. C'est l'étape d'alignement. Elle utilise comme modèle un génome humain de référence sur lequel le logiciel « aligne », de la manière la plus pertinente possible, tous les petits morceaux de 150 bases, en comparant une à une toutes les séquences d'ADN.



4 DÉTECTION DE VARIANTS

➜ Les algorithmes de calcul comparent, base par base, les séquences de l'ADN du patient par rapport au génome de référence. Toutes les anomalies (appelées variants) sont repérées : modification d'une base, délétion ou insertion d'une séquence d'ADN, etc.



5 ANNOTATIONS

➜ Chaque variant est annoté par l'attribution d'un score de gravité, allant de 1 à 5. Pour ce faire, les algorithmes de calcul se réfèrent à de multiples sources d'information : fréquence de la variation observée dans la population, bases de données des variations responsables des maladies rares, impact potentiel du changement d'une base sur la structure de la protéine codée par le gène, etc. La comparaison entre l'ADN des deux parents et celui de l'enfant malade apporte aussi, à cette étape, un éclairage supplémentaire. À noter que la plupart des variants sont non délétères, c'est-à-dire sans conséquences connues, ni biologiques, ni cliniques.

1. Non délétère
2. Probablement non délétère
3. Significations inconnue
4. Probablement délétère
5. Clairement délétère

6 POST-ANALYSE EN VUE DU DIAGNOSTIC

➜ Les fichiers de séquençage sont transmis à des généticiens et médecins, qui finalisent leur analyse pour aboutir, quand c'est possible, au diagnostic, c'est-à-dire à l'identification de la ou des anomalies génétiques responsables de la pathologie. Pour cela, ils disposent de leurs propres bases de données, liées à des structures familiales, et d'outils bio-informatiques de post-analyse. Dans un avenir proche, les progrès de la bio-informatique permettront d'élaborer des outils d'aide à la décision à l'usage du médecin : des modèles numériques prédictifs, basés sur l'analyse de milliers de profils de malades déjà traités ou en cours de traitement, proposeront au médecin la démarche thérapeutique la plus adaptée.

Les défis du CEA Décembre 2017 - Janvier 2018 **N°223**
Sylvie Rivière et Aurélien Boudault (infographie), avec **Jean-François Deleuze, François Sigaux, Vincent Meyer, CEA - DRF**

TOUT S'EXPLIQUE

© iStock

Et demain ? Vers le patient numérique

Un algorithme informatique pourra-t-il demain proposer à un médecin une stratégie thérapeutique pour son patient ? Oui, grâce au CAD, le « collecteur analyseur de données », prochainement mis en place dans le cadre du Plan France médecine génomique 2025. D'ici quelques années, chaque patient pourra se voir prescrire le séquençage complet de son génome. Ces informations génétiques seront recueillies au sein du CAD et complétées par l'ensemble de ses données cliniques (caractéristiques, parcours de soins, environnement, etc.). Objectif : constituer, à l'échelle nationale, des familles numériques de patients semblables, à la manière d'un médecin qui construit sa propre mémoire sur telle ou telle pathologie. Assorti de puissants algorithmes de traitement de données et de statistiques, le CAD sera un outil d'aide à la décision à l'usage du médecin : il lui proposera, grâce à des modèles numériques prédictifs, la démarche thérapeutique la plus adaptée à son patient. L'idée est de mettre en commun les données de tous pour aboutir à une prise en charge optimale de chaque malade.

La médecine du futur s'oriente ainsi vers une médecine de données. Aux données de génomique (apportées par le séquençage du génome) viendront s'ajouter celles de la protéomique (étude des protéines produites), de la métabolomique (suivi du métabolisme) et de la transcriptomique (processus de transcription du génome par les ARN messagers). C'est l'avènement du « patient numérique », une révolution qui devra prendre en compte les questions éthiques et de sécurité, la formation des professionnels de santé à ces nouveaux outils et la plus grande transparence sur le fonctionnement des algorithmes.

Le CEA, expert en génomique

Le CEA, *via* le CNRGH[1], est un acteur majeur de la génomique française. Très investi dans la mise en œuvre du Plan France médecine génomique 2025, il sera responsable, entre autres, du Crefix[2], le centre français de R&D de la médecine personnalisée.

Le CNRGH s'appuie sur une expertise de pointe et dispose d'importants moyens de recherche : biobanque d'échantillons, plateforme technologique de séquençage et de génotypage à haut et moyen débits, laboratoire d'épigénétique, laboratoire de bio-informatique assorti d'une puissante capacité de stockage et de traitement de données numériques… Il est impliqué dans la réalisation de projets ambitieux parmi lesquels le décryptage des causes génétiques des maladies rares, l'étude des bases génétiques de l'autisme ou encore les causes génétiques de la toxicité des traitements du cancer du sein.

1. Centre national de recherche en génomique humaine.
2. Centre de référence, d'innovation, d'expertise et de transfert.

les défis du cea 223

TOUT S'EXPLIQUE

Pipeline bio-informatique

Le terme « *pipeline* » est communément admis dans la communauté scientifique pour décrire une succession de traitements automatiques du signal au travers d'une chaîne de logiciels et d'algorithmes. Il s'emploie dans de nombreux domaines, notamment en génomique, pour l'analyse des gènes humains et de leurs fonctions.

ENJEUX

La recherche des maladies d'origine génétique (maladies rares, certains cancers, etc.) a fait des bonds de géant, grâce au formidable essor des technologies : séquençage à haut débit de l'ADN et traitement bio-informatique des données. Séquencer le génome d'un patient (c'est-à-dire déterminer la succession des bases de son ADN) nécessite aujourd'hui moins de quatre jours, et génère plus de 100 gigaoctets de données, soit l'équivalent en place mémoire de 200 heures de vidéo haute définition. Ces données sont traitées et analysées en des temps record, grâce à des *pipelines* bio-informatiques ultra-performants, qui trient, comparent et classent les fragments d'ADN, et à la puissance de supercalculateurs dépassant le pétaflops, c'est-à-dire capables de réaliser plus d'un million de milliards d'opérations par seconde. Ces prouesses technologiques ouvrent la voie à la médecine personnalisée, avec des traitements médicaux adaptés au profil génétique de chacun.