Direction de l'**E**valuation des **M**édicaments **E**t des produits **B**iologiques
Département de toxicologie
Chef de département : D. Masset
Secrétaire scientifique : A. Sanh

CONCEPT PAPER : LE SEQUENÇAGE A HAUT DEBIT METHODES ET ENJEUX EN MEDECINE, PHARMACOLOGIE ET TOXICOLOGIE

Concept Paper, Version 1, 8 décembre 2011

Document préparé par le Groupe de Travail « Innovation non Clinique »

- **Coordinateur de la rédaction**
  - S. LE CROM
- **Président**
  - J.-R. CLAUDE
- **Membres**
  - L. DOMENJOUD
  - E. FATTAL
  - J. GUILLEMAIN
  - A. GUILLOUZO
  - S. LE CROM
  - A. LE PAPE
  - S. LERONDEL
  - P. LESCUYER
  - B. MAILLERE
  - F. MOREL
  - M. PALLARDY
  - C. PINEAU
  - T. RABILLOUD
  - R. RAHMANI
- **Expert invité**
  - D. MARZIN
- **Représentants de l'Afssaps invités**
  - D. ABDON
  - D. SAUVAIRE

Le séquençage à haut débit : méthodes et enjeux en médecine, pharmacologie et toxicologie

Version 1

1. Introduction

Le séquençage de l'ADN par la méthode de synthèse enzymatique de Sanger (voir encadré ci-dessous), a été utilisé pendant plus de 30 ans pour lire le code génétique des organismes vivants. La réalisation la plus emblématique de cette approche est sans aucun doute le décryptage du génome humain, projet phare de la génétique à la fin des années 2000. La dernière génération des séquenceurs à capillaires, utilisant la technique Sanger, permet aujourd'hui de lire jusqu'à 2 millions de bases en une demi-journée. Cependant en 2007 sont apparus sur le marché des machines dotées de débits de 50 à 1 000 fois supérieurs. Ces séquenceurs de « nouvelle » génération ont permis de s'affranchir d'un certain nombre de biais de la méthode Sanger comme la nécessité de cloner l'ADN à séquencer. C'est grâce notamment à la lecture de plusieurs millions de séquences en parallèle que ces nouveaux séquenceurs à « haut débit » ont pu révolutionner les analyses en génomique (1).
Avec des coûts en baisse et un champ d'application très étendu, ces méthodes vont dans les années à venir complètement modifier la façon dont les scientifiques envisagent les études en génomique. En outre, la mise à disposition par les fabricants, d'appareils abordables et automatisés, va rendre ces méthodes incontournables.

Ce document à pour objectif de faire le point sur les différentes méthodes de séquençage à haut débit accessibles en France aujourd'hui en routine, de faire le tour des applications de ces outils dans le domaine de la médecine, de la pharmacologie, de la toxicologie et de voir quelles sont les limites et les difficultés actuelles de ces technologies.

Le séquençage de l'ADN par la méthode de Sanger

Le séquençage de l'ADN consiste à déterminer l'ordre d'enchaînement des nucléotides pour un fragment d'ADN donné. Deux méthodes ont été développées dans les années 70, l'une par l'équipe de Walter Gilbert, aux États-Unis, et l'autre par celle de Frederick Sanger, au Royaume-Uni. L'approche de Sanger est une méthode par synthèse enzymatique qui consiste à initier la polymérisation de l'ADN à l'aide d'un petit oligonucléotide (amorce) complémentaire à une partie du fragment d'ADN à séquencer. L'élongation de l'amorce est réalisée par une ADN polymérase dépourvue d'activité exonucléase en présence d'un mélange des quatre désoxy-ribonucléotides (dATP, dCTP, dGTP, dTTP) et une faible concentration de quatre didésoxynucléotides (ddATP, ddCTP, ddGTP ou ddTTP) chacun associé à un marqueur fluorescent différent. Une fois incorporés dans le nouveau brin synthétisé, ces didésoxynucléotides empêchent la poursuite de l'élongation. Il en résulte un mélange de fragments d'ADN de tailles croissantes, qui se terminent à toutes les positions dans la séquence. Ces fragments sont ensuite séparés par électrophorèse capillaire sur un gel de polyacrylamide ce qui permet ainsi de lire la suite de chacune des bases dans la séquence. Une dernière étape de traitement bioinformatique permet alors la reconstruction d'un génome entier à partir de tous les fragments séquencés.

2. Les techniques actuelles de séquençage à haut débit

C'est en 2007 qu'est apparue sur le marché la première génération des appareils de séquençage à haut débit. Globalement leurs technologies sont assez proches et fonctionnent en 3 étapes. La première consiste en la préparation et l'amplification des molécules d'ADN à analyser. La seconde permet l'incorporation des bases complémentaires du brin à séquencer. Enfin la dernière étape comprend la lecture de la séquence proprement dite. Trois technologies sont apparues de façon quasi simultanée : le pyroséquençage, le séquençage avec des terminateurs réversibles et le séquençage par ligation. Les caractéristiques de chacune de ces méthodes sont décrites dans le tableau 1 et détaillées ci-dessous.

a. Pyroséquencage (Roche)

Avec cette méthode, l'ADN subit après fragmentation une étape d'amplification par PCR. Cette PCR est réalisée en émulsion, ce qui permet d'effectuer plusieurs millions de réactions indépendantes dans un seul tube. Le mélange obtenu est ensuite déposé sur une plaque ou chaque emplacement ne peut contenir qu'une seule molécule d'ADN amplifiée. La réaction de séquençage par synthèse est alors initiée base par base. La lecture de chaque base incorporée est révélée à l'aide d'une réaction chemoluminescente et détectée par une caméra CCD.
Cette technologie permet d'obtenir aujourd'hui jusqu'à 1 million de séquences pouvant atteindre jusqu'à 400 bases. Les erreurs de séquences détectées sont majoritairement des insertions/délétions dues aux régions homopolymères (répétitions identiques de la même base).

b. Séquençage à l'aide de terminateurs réversibles (Illlumina)

Pour cette technique, l'amplification de l'échantillon à analyser ne s'effectue pas en solution mais sur un support solide. La réaction de séquençage est alors réalisée directement sur le support où l'ADN a été amplifié. Elle se déroule position après position en ajoutant un mélange contenant toutes les bases associées chacune à un fluorophore différent. L'extrémité de ces bases est protégée pour empêcher l'addition de bases supplémentaires à chaque cycle d'incorporation. Une lecture laser permet alors de détecter simultanément toutes les positions incorporées. Le clivage des fluorophores permet ensuite l'incorporation de la base suivante. La lecture est effectuée ainsi cycle après cycle.
Cette méthode permet l'acquisition en parallèle de plus de 3 milliards de séquences de 100 bases de long. Chaque position étant lue l'une après l'autre, les erreurs principales de cette technologie sont des erreurs de substitution d'une base par une autre.

c. Séquençage par ligation (Applied Biosystems)

La première étape d'amplification de la méthode de séquençage d'Applied Biosystems est identique à celle du pyroséquençage. Par contre, les séquences amplifiées sont fixées sur un support solide au lieu d'une plaque. La réaction de séquençage s'effectue ensuite par un système assez complexe de cycles de ligation et de clivage. Cette technique va permettre non seulement la lecture de la séquence mais inclut un système de correction des erreurs d'incorporations.
Grâce à cette méthode, il est possible de lire jusqu'à un milliard et demi de séquences en parallèle de 75 bases de long. Le système incorporé de correction d'erreurs associé à l'utilisation de la ligase rend cette technologie très fiable.

	Pyroséquençage	Terminateurs réversibles	Ligation
Fournisseur	Roche	Illumina	Applied Biosystems
Appareils (Nom commercial)	454 GS FLX 454 GS Junior	Genome Analyzer HiSeq MiSeq	SOLiD 5500XL
Longueur des lectures (pb)	400	2x 150	2x 75
Nombre de lectures	1 million	3 milliards	3 milliards
Données produites	600 Mb	600 Gb	600 Gb
Durée du séquençage	10 heures	11 jours	8 jours

Tableau 1 : Présentation des caractéristiques des différentes méthodes de séquençage accessibles en routine aujourd'hui en France.

3. Les applications du séquençage à haut débit

Le champ d'application de ces nouvelles méthodes de séquençage est très vaste. En effet à partir du moment où il est possible d'obtenir une molécule d'ADN, le séquençage à haut débit peut-être utilisé. Grâce à la capacité de ces machines à fournir de grandes quantités de séquences ou de travailler sur un grand nombre d'échantillons en parallèle, ces outils permettent de couvrir plusieurs technologies différentes employées jusqu'à maintenant comme les puces à ADN, le séquençage classique de Sanger ou la PCR quantitative à haut débit. Les applications disponibles avec le séquençage à haut débit peuvent globalement se regrouper en 3 grandes catégories : le séquençage *de novo*, le reséquençage et les analyses fonctionnelles.

a. Le séquençage *de novo*

La première application qui a été mise en avant grâce à ces nouvelles technologies de séquençage, ce sont leurs capacités à remplacer le séquençage Sanger traditionnel et donc à fournir la séquence de génomes inconnus. Pour réussir à obtenir des versions de génome de bonne qualité il est souvent nécessaire de combiner plusieurs méthodes. Le pyroséquençage permet grâce à des lectures longues de construire une première version du squelette du génome quand les méthodes par terminateurs réversibles vont corriger les erreurs présentes dans cette première reconstruction pour produire un brouillon du génome de qualité. Dans le domaine médical ces outils sont utilisés pour la découverte de génomes d'agents pathogènes inconnus ou de nouveaux virus (2).

b. Le reséquençage

Toujours dans le domaine où la connaissance de la séquence est importante, le reséquençage est utilisé quand la séquence du génome de référence est déjà connue. Le séquençage à haut débit est alors employé pour connaître quelles sont les variations génomiques de l'échantillon qui est étudié en comparaison avec celui pris comme référence. Ces approches sont certainement parmi les plus employées actuellement dans le domaine médical. Ces outils ont typiquement pour vocation de remplacer les méthodes traditionnelles d'hybridation génomique comparative (CGH) et de permettre de préciser les diagnostics soit de façon préventive soit pour caractériser une pathologie déjà déclarée. De nombreux articles ont été publiés dans ce domaine et le foisonnement des études d'associations en cours démontre de l'engouement réel du monde médical pour ces approches. Il est ainsi possible par exemple d'effectuer un diagnostic prénatal non invasif pour certaines maladies génétiques (3), de détecter les variants associés à

des maladies génétiques (4-8), de typer ou de suivre l'évolution des tumeurs cancéreuses chez les patients (9-11). Dans le domaine de la toxicologie, on peut imaginer que le séquençage permettra non seulement de suivre les effets mutagènes potentiels des molécules thérapeutiques sur les tissus qui ne sont pas les tissus cibles, mais on sera également capable de décrire précisément les modifications génomiques induites.
Le reséquençage peut également être utilisé pour caractériser les différentes souches d'agents pathogènes. Dans le cas d'études épidémiologiques, le séquençage à haut débit va permettre des gains de temps très importants par rapport aux méthodes traditionnelles (12), ce qui peut s'avérer crucial en situation d'urgence. L'exemple de la découverte de l'origine sud asiatique de la souche de choléra qui a infecté Haïti en 2010 grâce au séquençage à haut débit, a ainsi mis en évidence que la contamination provenait d'une région géographique éloignée probablement due aux activités humaines (13).
Enfin le dernier domaine d'application du séquençage à haut débit, à la frontière entre reséquençage et séquençage *de novo*, concerne la métagénomique. Le but est de découvrir dans un mélange complexe l'ensemble des organismes qui le composent comme par exemple la flore intestinale (14). On peut ainsi imaginer qu'en séquençant le sérum de patients, il sera possible de déterminer quels sont les agents pathogènes qui y sont présents et qui sont à l'origine de la pathologie observée. Il pourrait aussi être possible d'envisager en toxicologie le suivi de l'incorporation de vecteurs viraux dans différents tissus d'un organisme en dehors de ceux ciblés par le vecteur de thérapie génique.

c. Les analyses fonctionnelles

Le dernier champ d'application des méthodes de séquençage à haut débit concerne la génomique fonctionnelle. Dans ce domaine, l'important n'est pas de connaître la séquence d'ADN des échantillons, mais plutôt de quantifier le nombre et le type d'éléments biologiques présents. Il est par exemple possible de déterminer quelles quantités et quels types d'ARN s'expriment dans une cellule (RNA-Seq)(15), voire sélectivement les ARNs messagers fixés sur les ribosomes et donc en cours de traduction (traductome)(16), de détecter les phénomènes de transcription qui sont utilisés pour la régulation à travers les longs ARN non codants et toute la famille des petits ARN (Small RNA-Seq), de connaître les régions de l'ADN où se fixent les facteurs de transcription (ChIP-Seq)(17) et de déterminer les modifications épigénétiques d'un génome en cartographiant ses sites de méthylation (MeDIP-Seq)(18). Ces applications vont concurrencer de plus en plus les méthodes traditionnelles basées sur les puces à ADN. Le séquençage à haut débit permet en effet de s'affranchir des biais de l'hybridation, il dispose d'une gamme dynamique plus grande et la résolution des résultats obtenus est disponible à la base près. Ces outils sont donc intéressants pour déterminer les effets des molécules thérapeutiques lors de test *in vitro* sur des cultures de cellules, *in vivo* sur des modèles animaux ou lors des essais cliniques. Le fait de pouvoir soit être exhaustif sur l'ensemble de toutes les réponses génomiques possibles, soit de cibler une région précise et de passer un grand nombre d'échantillons en parallèle est un atout important pour la caractérisation et le criblage de biomarqueurs (19).

4. Limitations actuelles des méthodes de séquençage à haut débit

Les méthodes de séquençage à haut débit de première génération ont maintenant atteint la maturité technologique nécessaire à leur adoption large par la communauté scientifique et médicale. Les fabricants de matériels ne s'y sont pas trompés, et lancent d'ailleurs en 2011 chacun des versions de leurs matériels dédiés aux applications de diagnostic. C'est le cas de Roche avec le GS Junior, d'Illumina avec le MiSeq et d'Applied Biosystems avec le Ion Personal Genome Machine. Dans le même temps la seconde génération d'appareils commence à arriver sur le marché. Elle ambitionne de résoudre deux des limitations actuelles des premières générations de machines à haut débit. D'une part la nécessité d'amplifier l'échantillon à analyser, les machines de seconde génération seront capables de lire directement la séquence d'ADN. D'autre part le temps d'acquisition des appareils actuels est assez lent car il se fait par cycle. Les nouveaux séquenceurs seront eux capables d'effectuer la lecture de l'ADN en temps réel.

Cependant face à cette escalade technologique et à l'ébullition autour des applications potentielles, il faut garder à l'esprit que le séquençage à haut débit doit faire face encore à plusieurs limitations. Tout d'abord, c'est une technologie encore jeune. Il reste de nombreuses zones d'ombre en ce qui concerne ces approches. Les biais techniques ne sont pour le moment qu'au début de leur caractérisation. On manque encore de recul sur les capacités de ces outils à quantifier précisément les événements biologiques complexes (épissage alternatif, site de fixation à l'ADN des facteurs de transcription…). Ensuite, le séquençage à haut débit produit des quantités très importantes de données dont la gestion n'est pas totalement maîtrisée et encore moins automatisée. Les résultats générés se chiffrant aujourd'hui en téraoctets, la question de savoir quoi garder et pendant combien de temps se pose de façon cruciale. Ce point est d'autant plus important que les études sur les effets de molécules thérapeutiques ont lieu pendant plusieurs années et qu'elles nécessitent un grand nombre d'expériences. Des questions importantes se posent également sur le traitement des résultats obtenus. Si par exemple dans le cas des puces à ADN on dispose de plus de 15 ans de recul technologique et de méthodologies statistiques bien déterminées pour corriger les données brutes et obtenir des résultats fiables, ce n'est pas du tout le cas en ce qui concerne les résultats du séquençage à haut débit. Les questions de normalisation ne sont pas tranchées, et les modèles statistiques à appliquer aux données ne sont toujours pas déterminés sans ambiguïtés. Pour terminer, dans le domaine du séquençage du génome des patients, des questions éthiques importantes se posent. Il est maintenant possible d'accéder au génome complet d'un individu. Tous les aspects éthiques de ces outils doivent donc être envisagés et surtout débattus pour y offrir une réponse appropriée.

5. Conclusion

On vient de le voir, les méthodes de séquençage à haut débit sont porteuses de grands espoirs en terme d'innovations médicales, pharmacologiques et toxicologiques. Cependant, les difficultés soulevées par l'analyse et le traitement des données produites sont quasiment aussi importantes et nécessitent une attention particulière de tous les décideurs dans le domaine de la santé. Il paraît impératif pour permettre l'adoption de ces méthodes dans le domaine du suivi toxicologique de pouvoir répondre dans le futur aux questions suivantes. Comment assurer la reproductibilité et la standardisation des méthodes d'études face aux évolutions incessantes des technologies ? Quels sont les

protocoles d'analyses des données et les contrôles à employer pour garantir la fiabilité des résultats proposés ? Enfin, de quelle manière le stockage et la pérennisation des données seront-t-ils assurés ?

6. Références

1. Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31-46.
2. Gaynor, A.M., Nissen, M.D., Whiley, D.M., Mackay, I.M., Lambert, S.B., Wu, G., Brennan, D.C., Storch, G.A., Sloots, T.P. and Wang, D. (2007) Identification of a novel polyomavirus from patients with acute respiratory tract infections. *PLoS Pathog*, **3**, e64.
3. Chiu, R.W., Chan, K.C., Gao, Y., Lau, V.Y., Zheng, W., Leung, T.Y., Foo, C.H., Xie, B., Tsui, N.B., Lun, F.M. *et al.* (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci U S A*, **105**, 20458-20463.
4. Cirulli, E.T. and Goldstein, D.B. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*, **11**, 415-425.
5. Isidor, B., Lindenbaum, P., Pichon, O., Bezieau, S., Dina, C., Jacquemont, S., Martin-Coignard, D., Thauvin-Robinet, C., Le Merrer, M., Mandel, J.L. *et al.* (2011) Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis. *Nat Genet*, **43**, 306-308.
6. Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A. *et al.* (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*, **362**, 1181-1191.
7. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, **42**, 30-35.
8. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C. *et al.* (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet*, **43**, 585-589.
9. Kan, Z., Jaiswal, B.S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H.M., Yue, P., Haverty, P.M., Bourgon, R., Zheng, J. *et al.* (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*, **466**, 869-873.
10. Meyerson, M., Gabriel, S. and Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, **11**, 685-696.
11. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D. *et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature*, **472**, 90-94.
12. Harris, S.R., Feil, E.J., Holden, M.T., Quail, M.A., Nickerson, E.K., Chantratita, N., Gardete, S., Tavares, A., Day, N., Lindsay, J.A. *et al.* (2010) Evolution of MRSA during hospital transmission and intercontinental spread. *Science*, **327**, 469-474.
13. Chin, C.S., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., Jean-Charles, R.R., Bullard, J., Webster, D.R., Kasarskis, A., Peluso, P. *et al.* (2011) The origin of the Haitian cholera outbreak strain. *N Engl J Med*, **364**, 33-42.
14. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59-65.
15. Ozsolak, F. and Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, **12**, 87-98.
16. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. and Weissman, J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218-223.
17. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, **10**, 669-680.
18. Laird, P.W. (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet*, **11**, 191-203.
19. Schwarzenbach, H., Hoon, D.S. and Pantel, K. (2011) Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer*, **11**, 426-437.