IDENTIFICATION DE NOUVEAUX AGENTS PATHOGÈNES: LA RÉVOLUTION DU SÉQUENÇAGE HAUT DÉBIT

par Marc Eloit[1]

Le diagnostic microbiologique souffre encore de graves lacunes. En effet, la routine des techniques microbiologiques effectuées en microbiologie clinique est souvent incapable d'identifier les agents pathogènes responsables (par exemple, ils donnent des résultats négatifs dans plus de 40% des cas d'infections aiguës respiratoires profondes et dans plus de 50% des cas d'encéphalite aiguë). L'utilisation d'outils modernes pour le diagnostic sans *a priori* pourrait rapidement améliorer nos connaissances sur l'étiologie de certains syndromes. En plus des techniques basées sur l'amplification et le séquençage de l'ARNr procaryotes et eucaryotes, des techniques à haut débit ont été mises au point, basée sur l'amplification aléatoire des génomes et / ou ARNm suivie de l'identification de séquences pertinentes par :
– (I) l'hybridation sur des puces à faible densité ou à haute densité en séquences virales, bactériennes, fungiques ou parasitaires (Virochip, Wang *et al.*, 2003) Greenechip (Palacios *et al.*, 2007). Certaines de ces puces permettent un re-séquençage et ont donc la propriété de détecter des séquences inconnues, mais ces séquences doivent rester proches (> 90%) de celles présentes sur la puce. La puce développée en collaboration entre l'Institut Pasteur et Affymetrix (Berthet et al, 2008) est un des exemples de ces approches,
– (Ii) séquençage haut débit ou *High Throughput Sequencing* –HTS- (pyroséquençage ou Illumina) suivi du filtrage des séquences obtenues par l'élimination des séquences de l'hôte (humaines ou animales). Afin de limiter le nombre de séquences à analyser, et d'éviter la contamination par de l'ADN / ARN cellulaires, ces techniques ont d'abord été utilisées après enrichissement et purification des génomes d'agents infectieux par des processus physiques. Par exemple, cette méthodologie a permis d'identifier un nouveau parvovirus humain (bocavirus) dans des échantillons respiratoires humains (Allander *et al.*, 2005). Plus récemment, grâce au pyroséquençage, l'approche a été appliquée au séquençage de la totalité de l'ARN, le filtrage étant effectué a posteriori sur les bases de données. Comme souligné par le laboratoire de Lipkin (Quan *et al.*, 2008), cela se justifie par le fait que l'infection « *active should be associated with transcription* ». De cette façon, l'agent causal probable du syndrome de dépérissement des colonies d'abeilles a été identifié (Cox-Foster *et al.*, 2007), ainsi qu'un nouvel arénavirus, responsable d'un groupe de maladies mortelles après transplantation chez des patients greffés par des organes venant du même donneur infecté (Palacios *et al.*, 2008), et un polyomavirus nouveau responsable de carcinome à cellules de Merkel (Feng *et al.*, 2008). Cette technique peut également être utilisée pour identifier les génomes d'agents infectieux, lorsqu'ils sont présents à des concentrations élevées par rapport au bruit de fond de l'ADN génomique de l'hôte. De cette façon, il a été possible d'identifier chez 12 enfants diarrhéiques neuf « nouveaux » virus, dont sept sous-types ou nouveaux génotypes de picobirnavirus, entérovirus, virus TT ou norovirus, avec deux nouvelles espèces de nodavirus et d'astrovirus (Finkbeiner *et al.*, 2009). Des bactéries pathogènes ont également été identifiés dans des échantillons de selles de patients par séquençage profond de l'ADN bactérien (Nakamura *et al.*, 2008), et l'analyse de la complexité de la flore bactérienne de l'intestin dans les situations pathologiques a été réalisée par HTS d'ARN 16S (Zhang *et al*, 2009).

L'Institut Pasteur a récemment (début 2009) mis en place un tel programme de découverte d'agents pathogènes humains et des animaux (*Pathodisc pour Pathogen Discovery*), basé sur l'interface de services hospitaliers avec un réseau de plates-formes coordonnées et unités de recherche à l'intérieur de l'Institut Pasteur, et dont l'objectif est d'identifier et caractériser des agents infectieux humains ou animaux, soit nouveaux soit inconnus pour certaines étiologies. *Pathodisc* est conçu comme un « pipeline » permettant de traiter des échantillons issus du domaine humain ou animal. Il s'appuie également sur les ressources en matière d'anatomie pathologique du campus et de bioinformatique. À partir de 2011, une startup

[1] Institut Pasteur, responsable du programme PathoDisc et UMR 1161 ANSES-INRA-ENVA, École nationale vétérinaire d'Alfort (marc.eloit@pasteur.fr).

(Pathoquest), spin-off de l'IP et de l'ENVA a été créée et a renforcé les moyens des programmes, tout en vendant des services correspondant à l'industrie.

La validation des techniques de séquençage Illumina et pyroséquençage a été conduite, ainsi que la construction d'un pipeline permettant la gestion bioinformatique des données brutes de séquence. Il consiste à trier les flux de séquences générées par le séquenceur, en séparant les séquences d'intérêt (agents candidats) de la majeure partie des séquences de l'hôte. Bien que l'analyse informatique des données montre certaines similitudes avec l'analyse des données de métagénomique, la petite quantité prévue de séquences d'intérêt, ainsi que les caractéristiques des séquences Illumina (courtes et sujettes à erreur), nécessitent le développement d'approches spécifiques.

Les grandes lignes de l'analyse sont les suivantes : (i) nettoyage des séquences restantes pour faciliter leur analyse ultérieure et identification (par exemple en éliminant les nucléotides de faible qualité et en réduisant la redondance par un processus d'assemblage), (ii) filtrage des séquences humaines, (iii) en utilisant des approches différentes, réalisation d'une assignation taxonomique. Ces méthodes sont explicitées ci-dessous :

La grande majorité des séquences proviendront de l'hôte. Il est donc nécessaire d'éliminer ces séquences par une procédure de comparaison de base de données soustractive. À cette fin, toute la séquence du génome de l'homme et de différentes espèces a été sélectionnée et analysée par les logiciels SOAP et BlastN. Les meilleurs paramètres à utiliser ont été déterminés au cours de la phase pilote. Afin d'optimiser l'analyse subséquente des séquences, les séquences non hôtes font l'objet d'une procédure de regroupement (« *contigs* »), si possible. Une certain nombre d'assembleurs (CLC, SOAP, Velvet) ont été testés pour leur efficacité sur des données hétérogènes et sont maintenant utilisées dans notre pipeline de traitement de l'information. La comparaison des séquences non hôtes avec des séquences déjà connues est particulièrement efficace car elle s'appuie sur un dispositif de recherche de séquences BlastN et BlastX sur des bases de données virales maintenues localement, en utilisant Paracel, un dispositif de distribution des taches de Blast sur différents cœurs de serveur. L'étape finale du processus d'analyse de la séquence consiste en une assignation provisoire taxonomiques et d'une étude phylogénétique des séquences ayant une e value considérée comme significative (de l'ordre de - 5).

L'exécution d'une étude pilote (en utilisant des échantillons contaminés volontairement), a permis de déterminer la sensibilité et la spécificité de la technique. Par exemple, nous avons montré que nous étions capables de détecter dans une étude à l'aveugle un niveau de contamination compris entre 1 et 10^3 pfu ou cfu / ml de virus différents (rage, poliovirus, virus d'Epstein Barr, Hepatitis B, Adenovirus, Cytomegalovirus, Parvovirus B19, Influenzavirus A, Influenzavirus B, Bovine viral Diarrhea virus, virus de l'Hepatite C) et des bactéries (Klebsiella sp.). Aucun résultat faussement positif n'a été obtenu. Ces résultats étaient similaires, voire plus sensibles que des PCR ciblées correspondantes, sans avoir besoin de connaissances préalables de l'agent infectieux cible cible (Cheval *et al.,* 2011). Ce niveau de sensibilité a été atteint en générant de 5-7 millions de séquences à partir d'un échantillon (soit une profondeur de séquençage bien inférieure à celle désormais utilisée , de l'ordre de 100-150 millions).

Nous avons identifié différents nouveaux virus humains : un nouveau polyomavirus (Sauvage et al, 2011a), le premier gyrovirus de mammifère (Sauvage *et al.,* 2011b), différents papillomavirus au sein de viromes cutanés complexes (Foulongne *et al.,* 2012). Nous avons également identifié plusieurs nouveaux virus chez le porc, dont l'un, proche des parechovirus humains, définissant un nouveau genre au sein des Picornaviridae (Sauvage *et al.,* 2012) et d'autres en cours d'analyse. Une souche neurovirulente de virus Theiler a été décrite chez la souris (Buckwalter *et al.,* 2011) Au cours d'analyses conduites chez des tiques collectées en Alsace, nous avons également identifié plusieurs virus de familles comportant des membres pathogènes pour l'homme et l'animal. Enfin, nous avons utilisé cette approche pour identifier si des mutations étaient associées à la transmission interspécifique homme-porc du virus zoonotique de l'hépatite E (Bouquet *et al.,* 2012).

Actuellement, les échantillons disponibles, correspondant à différentes conditions médicales difficiles, y compris les encéphalites et vascularites et syndrome infectieux indéterminés chez des enfants souffrant d'un immunodéficit congénital (service Alain Fischer, Necker), des tumeurs, des échantillons issus de maladies infectieuses en Asie, ainsi que différents syndromes animaux sont également analysés

Les postulats de Koch classique ne peuvent être satisfaits pour des agents non cultivables, et la révision des postulats de Koch a été proposée, comme récemment discuté en profondeur (Lipkin, 2009). L'hypothèse de la causalité des séquences identifiées est conduite de la manière suivante : nous utilisons un critère d'association pour renforcer l'association de cause à effet dans une deuxième phase. Nous mettrons au

point des tests PCR basés sur les séquences identifiées, ou des séquences consensus pour les agents correspondants. Selon les résultats, lorsque l'évaluation de la charge de la séquence apparaît utile pour l'interprétation, des qPCR spécifiques sont développées. Nous recherchons si les agents identifiés sont plus fréquents dans les cas détectés que dans des contrôles (Fredericks et Relman, 2009), ou dans des échantillons biologiques non pertinents du même malade sélectionnés de manière adéquate.

La puissance de ces techniques apparaît désormais incontournable pour la découverte de pathogènes et est également développée pour le contrôle de produits biologiques. A court terme, les applications diagnostiques verront également probablement le jour.

RÉFÉRENCES BIBLIOGRAPHIQUES

(1) ALLANDER T., TAMMI MT., ERIKSSON M., *et al.*, 2005. – Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proc Natl Acad Sci U S A.* **102**, 12891-12896.

(2) BERTHET N., REINHARDT AK., LECLERCQ I., *et al.*, 2008. – Phi29 polymerase based random amplification of viral RNA as an alternative to random RT-PCR. *BMC Mol Biol.*, 9-77.

(3) BOUQUET J., CHEVAL J., ROGÉE S., PAVIO N., ELOIT M., 2012 Mar 28. – Identical Consensus Sequence and Conserved Genomic Polymorphism of Hepatitis E Virus during Controlled Interspecies Transmission .J Virol.. [Epub ahead of print].

(4) BUCKWALTER MR., NGA PT., GOUILH MA., FIETTE L., BUREAU JF., LAIRD ME., BUCHRIESER J., OZDEN S., CHEVAL J., ELOIT M., MANUGUERRA JC., GESSAIN A., BREY PT., FONTANET A., ALBERT ML., 2011 Jul. – Identification of a novel neuropathogenic Theiler's murine encephalomyelitis virus. J Virol., **85**(14), 6893-905. Epub 2011 May 4.

(5) CHEVAL J., SAUVAGE V., FRANGEUL L., DACHEUX L., GUIGON G., DUMEY N., PARIENTE K., ROUSSEAUX C., DORANGE F., BERTHET N., BRISSE S., MOSZER I., BOURHY H., MANUGUERRA CJ., LECUIT M., BURGUIERE A., CARO V, ELOIT M., 2011 Sep. – Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples.J Clin Microbiol.; **49**(9), 3268-75. Epub 2011 Jun 29.

(6) COX-FOSTER DL., CONLAN S., HOLMES EC., *et al.*, 2007. – A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*; **318**, 283-287.

(7) FENG H., SHUDA M., CHANG Y., MOORE PS., 2008. – Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science*, **319**, 1096-1100.

(8) FINKBEINER SR., ALLRED AF., TARR PI., *et al.*, 2008. – Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog.*, **4**, e1000011.

(9) FOULONGNE V. , SAUVAGE V., HEBERT C., DEREURE O., CHEVAL J., AR GOUILH M., PARIENTE K., SEGONDY M., BURGUIÈRE A., MANUGUERRA JC., CARO V., ELOIT M., – Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing, Plos One, in press.

(10) FREDERICKS DN., RELMAN DA., 1996. – Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin Microbiol Rev.*, **9**, 18-33.

(11) LIPKIN WI., 2009. –Microbe hunting in the 21st century. *Proc Natl Acad Sci U S A*, **106**, 6-7.

(12) NAKAMURA S., MAEDA N., MIRON IM., YOH M., IZUTSU K., KATAOKA C., HONDA T., YASUNAGA T., NAKAYA T., KAWAI J., HAYASHIZAKI Y., HORII T., IIDA T., 2008. – Metagenomic diagnosis of bacterial infections.Emerg Infect Dis., **14**(11), 1784-6.

(13) PALACIOS G., QUAN PL., JABADO OJ., *et al.*, 2007. – Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg Infect Dis.*,**13**, 73-81.

(14) PALACIOS G., DRUCE J., DU L., *et al.*, 2008. – A new arenavirus in a cluster of fatal transplant-associated diseases. *N Engl J Med.*, **358**, 991-998.

(15) SAUVAGE V., FOULONGNE V., CHEVAL J., AR GOUILH M., PARIENTE K., DEREURE O., MANUGUERRA JC., RICHARDSON J., LECUIT M., BURGUIÈRE A., CARO V., ELOIT M., 2011. – Human polyomavirus related to African green monkey lymphotropic polyomavirus.Emerg Infect Dis., **17**(8), 1364-70.

(16) SAUVAGE V., CHEVAL J., FOULONGNE V., GOUILH MA., PARIENTE K., MANUGUERRA JC. RICHARDSON J., DEREURE O. LECUIT M., BURGUIERE A., CARO V., ELOIT M., 2011. –

Identification of the first human gyrovirus, a virus related to chicken anemia virus. J Virol., **85**(15), 7948-50. Epub 2011 Jun 1.

(17) SAUVAGE V., AR GOUILH M., CHEVAL J., MUTH E., PARIENTE K., BURGUIERE A., CARO V., MANUGUERRA JC., ELOIT M., – A member of a new Picornaviridae genus is shed in pig feces, J. Virol, in press.

(18) QUAN PL., BRIESE T., PALACIOS G., IAN LIPKIN W., 2008. – Rapid sequence-based diagnosis of viral infection. *Antiviral Res.*, **79**, 1-5.

(19) WANG D., URISMAN A., LIU YT., *et al*., 2003. – Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.*, **1**, E2.

(20) ZHANG H., DIBAISE JK., ZUCCOLO A., *et al.*, 2009. – Human gut microbiota in obesity and after gastric bypass. *Proc Natl Acad Sci U S A*, **106**, 2365-2370.