

# Utilisation des phylogénies d'infections en épidémiologie

Samuel Alizon, Emma Saulnier

► **To cite this version:**

Samuel Alizon, Emma Saulnier. Utilisation des phylogénies d'infections en épidémiologie. Bulletin de veille scientifique Santé Environnement Travail de l'ANSES, 2017, pp.42-43. <<http://bvs.mag.anses.fr/>>. <anses-01533229>

**HAL Id: anses-01533229**

**<https://hal-anses.archives-ouvertes.fr/anses-01533229>**

Submitted on 6 Jun 2017

**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.

# Utilisation des phylogénies d'infections en épidémiologie

Samuel ALIZON et Emma SAULNIER | samuel.alizon@cnrs.fr

CNRS - Laboratoire Mivegec (UMR CNRS, IRD, UM 5290) - Montpellier

Mots clés : Bactérie, épidémiologie, évolution, infection, mutation, phylodynamique, séquençage, virus

Les virus évoluent rapidement. Ceci explique en partie les difficultés à mettre au point des vaccins ou des traitements (1). L'évolution est dite « neutre » quand les mutations génétiques n'affectent pas les traits de l'infection. La phylodynamique propose d'utiliser cette évolution neutre afin d'inférer des paramètres épidémiologiques (2,3). Il s'agit d'étudier la manière dont les virus se propagent, ce qui laisserait des traces dans leur génome. À la base de cette discipline, il y a le parallèle entre une phylogénie d'infection (représentation de l'évolution du virus au cours d'une épidémie) et une chaîne de transmission. En effet, chaque embranchement dans la phylogénie peut être vu comme une transmission et chaque feuille comme une fin d'infection. Ceci a conduit les biologistes de l'évolution à appliquer des modèles phylogénétiques au cas des phylogénies d'infections, par exemple pour approximer la prévalence des infections (4). Ces dernières années ont vu des progrès dans le développement de modèles statistiques permettant d'inférer des paramètres tels que le taux de reproduction de base (ou  $R_0$ ), qui caractérise la vitesse de propagation d'un pathogène. Ces méthodes ont en commun de reposer sur des modèles mathématiques qui calculent la probabilité d'observer une phylogénie pour un modèle sous-jacent donné et un jeu de paramètres donné (5,6). Des approches dites de phylogéographie permettent aussi de retracer la propagation d'épidémies dans l'espace, par exemple la grippe aviaire en Asie (7). Enfin, de nouvelles approches visent à maximiser la précision des modèles en combinant des données d'épidémiologie « classiques » (prévalence) et données génétiques (8). Les trois publications illustrent l'apport de la phylodynamique pour l'étude d'infections émergentes (Zika), sa faisabilité dans un contexte difficile (Ebola en Afrique de l'Ouest) et les défis soulevés par l'accumulation de séquences (le cas du virus du syndrome respiratoire et reproductif porcin).

## Le virus Zika dans les Amériques : premiers résultats épidémiologiques et génétiques

Faria NR, et al. (2016). Zika virus in the Americas: Early epidemiological and genetic findings. *Science*, vol. 352: p.345-9.

### Résumé

Le virus Zika appartient au genre des Flavivirus (comme la dengue). Il est généralement transmis à l'homme par un moustique du genre *Aedes* et peut causer des symptômes modérés (maux de tête, courbatures) et avoir des conséquences graves chez le fœtus. Zika a été découvert en 1947 en Ouganda. On en connaît deux génotypes (africain et asiatique). Entre 2007 et 2014, Zika a causé plusieurs épidémies dans le Pacifique. En mai 2015, il a été détecté au Brésil et, fin janvier 2016, 30.000 cas avaient été recensés (ce chiffre sous-estime fortement l'incidence car la majorité des cas sont asymptomatiques).

Afin d'en savoir plus sur l'origine de l'épidémie, les auteurs ont séquencé le génome complet de Zika chez sept individus. Ces sept séquences étaient peu différentes des séquences existantes. En inférant un arbre phylogénétique daté à partir de ces sept séquences et de 16 autres disponibles dans la base de données de référence, ils ont constaté que les séquences d'Amérique du Sud appartenaient au génotype asiatique et partageaient un ancêtre commun avec la souche de 2013 en provenance de Polynésie française. De plus, les séquences du Brésil sont éparpillées dans le groupe des séquences américaines, suggérant une propagation du Zika entre les pays. Toutefois, comme le Zika a d'abord été identifié au Brésil et, comme sa diversité génétique y est la plus élevée, les auteurs pensent que c'est là que l'épidémie Sud-Américaine a commencé.

Enfin, selon les auteurs, l'ancêtre commun aux séquences américaines se situerait entre août 2013 et avril 2014. Celui des séquences américaines et de Polynésie française remonterait à une date entre décembre 2012 et septembre 2013. Ceci suggère que le virus n'a pas été introduit au Brésil pendant la coupe du monde de football (été 2014) mais peut-être pendant la coupe des confédérations en 2013. Enfin, les répercussions des mutations de la souche Brésilienne de Zika sur les propriétés physico-chimiques des protéines du virus ne semblent pas avoir d'impact sur la transmission de ce dernier.

### Commentaire

Cet article montre que la phylodynamique est maintenant utilisée en routine pour analyser une « flambée épidémique ». L'analyse de quelques génomes permet d'en savoir plus sur l'origine de l'épidémie, en particulier dans le cas d'un pathogène mal connu. Ces méthodes permettent aussi de dater les événements, par exemple l'arrivée de Zika au Brésil. Le peu de séquences utilisées ici fait que l'incertitude autour des dates estimées est grande. C'est d'ailleurs une des limites de cette étude, surtout à l'heure où le séquençage est devenu facile même en conditions extrêmes comme l'illustre la seconde étude.

## Séquençage portable des génomes en temps réel pour la surveillance d'Ebola

Quick J, et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, vol. 530: p.228-32.

## Résumé

Le MinION est un système de séquençage très léger, et connectable sur ordinateur portable. Cette nouvelle technologie permet de séquencer rapidement de grandes séquences d'ADN (>50 kb). Les auteurs décrivent un protocole de séquençage et d'analyse de séquences rapide, basé sur la technologie MinION. Pour prouver sa praticité, ils l'ont utilisé pendant l'épidémie d'Ebola en Afrique de l'ouest. Au préalable, ils ont validé leur approche en la comparant au séquençage Illumina classique. Ils ont ensuite expédié l'ensemble de leur système en Guinée. En six mois, ils ont obtenu 142 séquences complètes sur 148 protocoles de séquençage par MinION. Celles-ci ont été combinées à 603 autres séquences d'Ebola récoltées depuis le début de l'épidémie en 2014 pour inférer un arbre phylogénétique daté avec le logiciel BEAST (*Bayesian Evolutionary Analysis Sampling Trees*). Dès les 10 premiers jours de séquençage, ils ont identifié que les souches persistantes en Guinée appartenaient à deux principales lignées établies dès le début de l'épidémie : GN1 qui est largement restée confinée en Guinée et SL3 qui a d'abord été détectée en Sierra Leone puis en Guinée vers la fin 2014. Grâce à leur jeu de données combinées, ils ont pu constater que GN1 et SL3 ont été toutes les deux détectées en Sierra Leone, début 2015. Ceci suggère qu'il y a eu des transmissions entre les deux pays.

## Commentaire

Cet article illustre la faisabilité des approches de phylodynamique. Même dans le cas d'un parasite extrêmement virulent sévissant dans des régions du monde particulièrement mal équipées, les techniques modernes permettent de séquencer des génomes entiers en temps réel. Toutefois, au-delà de ce « tour de force », les résultats biologiques décrits sont limités. En effet, ils apportent peu de nouveaux éléments par rapport aux études existantes sur l'origine et même sur la propagation de l'épidémie d'Ebola.

De plus, du fait qu'il séquence des molécules uniques d'ADN, le MinION est reconnu pour avoir un taux d'erreur non négligeable (de 10 à 20 %). Pour palier à ce défaut, les auteurs ont donc effectué plusieurs lectures des séquences (au moins 25 fois dans l'étude).

## Méthodes phylodynamiques dans une épidémie récente du virus du syndrome reproducteur et respiratoire porcin

Alkhamis MA. et al. (2016). Applications of Bayesian Phylogenetic Methods in a Recent U.S. Porcine Reproductive and Respiratory Syndrome Virus Outbreak. *Front Microbiol*, vol.7 : p.67-77

## Résumé

Le syndrome respiratoire et reproductif porcin (PRRS), endémique aux États-Unis, est causé par le virus du PRRS (PRRSV), qui se transmet très rapidement. L'épidémie de PRRSV est difficile à contrôler et de nouvelles souches de ce virus émergent, classées par des marqueurs RFLP (*Restriction Length Fragment Polymorphism*). Cette étude utilise des modèles phylodynamiques pour caractériser la propagation d'une souche virulente du PRRSV entre 2014 et 2015 aux États-Unis afin d'orienter les mesures de contrôle. Les

auteurs ont obtenu 6774 séquences ADN d'une protéine du PRRSV, échantillonnées entre Janvier 1998 et Avril 2015 à partir de cinq lieux de production porcine indépendants (identifiés A, B, C D et E) et qui pratiquent 3 types d'élevage (troues en naissance-sevrage, troues en naissance-engraissement, et cochons).

Une première analyse phylogénétique de l'ensemble des séquences leur a permis d'identifier un groupe de 288 séquences phylogénétiquement proches (majoritairement du type RFLP 1-7-4), échantillonnées entre Septembre 2003 et Mars 2015. Ce groupe a fait l'objet d'analyses épidémiologiques plus approfondies par le logiciel BEAST (*Bayesian Evolutionary Analysis Sampling Trees*). Les auteurs ont ainsi constaté que le nombre de porcs infectés par le PRRSV de type RFLP 1-7-4 était en rapide expansion. Ils ont également estimé que l'ancêtre commun de ce groupe de séquences remonte à Septembre 1996, date qui coïncide avec celle de l'ancêtre commun des séquences provenant du lieu C. D'autres analyses sur la dispersion du virus ont permis d'identifier le lieu C comme unique source des épidémies dans les lieux A, B, D et E. Selon l'étude, la dispersion depuis le lieu C aurait commencé en Septembre 2000. Enfin, en croisant la phylogénie avec les types d'élevage et lieu des localisations, les auteurs ont estimé que le pathogène circule uniquement depuis les fermes de troues vers les autres fermes et que l'épidémie aurait débuté dans les fermes de troues en Septembre 1999.

## Commentaire

L'article illustre la quantité de données analysables (des milliers de séquences et des données de localisation) mais aussi les limites des méthodes actuelles. Car, pour identifier les routes de dispersion du virus, les auteurs n'ont pu garder que 4 % des séquences. Quel serait l'impact des 96% restants ? De plus, les données n'ont été obtenues qu'à partir d'éleveurs volontaires, ce qui peut biaiser les résultats. Enfin, l'article illustre l'importance d'avoir une bonne connaissance du sujet d'étude afin de bien calibrer les modèles (par exemple les types d'élevages)

**CONCLUSION GÉNÉRALE**

Les avancées de séquençage sont en train de bouleverser l'épidémiologie. Comme les épidémies récentes d'Ebola et de Zika l'illustrent, il est possible d'avoir une couverture quasi-exhaustive de la diversité génétique des virus en temps réel. Les approches de phylogénétique permettent alors d'en savoir plus sur la propagation et la biologie des infections (taux de transmission, durée d'infection, origine géographique) et d'ajuster en temps réel les mesures de santé publique (campagnes de sensibilisation, dépistage et traitement, quarantaines). Ces approches sont, en effet, moins sensibles aux biais d'échantillonnage que les approches classiques. Deux défis méthodologiques se posent alors : d'une part, combiner plusieurs types de données (séquences, dates, prévalences, localisations), et d'autre part, utiliser de grands jeux de données, les approches actuelles fonctionnant essentiellement pour des modèles simples et des phylogénies limitées.

**GENERAL CONCLUSION**

*Progresses in sequencing are revolutionizing epidemiology. As illustrated by the recent Ebola and Zika outbreaks, it is now possible to get a good coverage of the virus genetic diversity in real-time. Phylodynamic approaches then provide information on the spread and the biology of infections (transmission rate, infection duration, geographical origin) and enable to adjust the public health measures (awareness, screening and treatment campaigns, quarantines) in real time. Indeed, those approaches are less sensitive to sampling bias than classical approaches. Two methodological challenges remain open: first, being able to combine several types of data (sequences, dates, prevalences, locations), and, second, using large datasets since current methods only work for simple models and small phylogenies (a few hundred leaves).*

**Lexique**

**Ancêtre commun** : En phylogénie, l'ancêtre commun d'un groupe d'individus est l'individu le plus proche en remontant dans le temps, et dont descendent tous les individus en question. En épidémiologie, on parle aussi de « patient zéro ».

**BEAST** (*Bayesian Evolutionary Analysis Sampling Trees*) : logiciel d'analyse bayésienne d'arbres phylogénétique qui permet d'effectuer des analyses phylodynamiques.

**Diversité génétique** : degré de variété des séquences au sein d'une même espèce.

**Phylogénie** : Arbre qui représente la distance et la relation de parenté entre des individus (typiquement représentés

par des séquences génétiques ou protéiques). Chaque feuille de l'arbre correspond à une séquence et chaque nœud d'embranchement à l'ancêtre commun des feuilles qui lui succèdent.

**Phylogéographie** : approche consistant à combiner données de séquences, date de prélèvement et lieu de prélèvement afin d'inférer l'histoire de propagation d'un organisme.

**$R_0$**  : taux de reproduction de base.

**RLFP** (*Restriction Length Fragment Polymorphism*) : polymorphisme de longueurs de fragments de restriction, parfois utilisé pour caractériser la souche d'un pathogène. Le génome d'un pathogène peut être coupé en fragments par une enzyme de restriction (protéine). Le profil des découpes peut varier d'un génome à un autre et peut donc être utilisé pour reconnaître un groupe de pathogènes appartenant à une même souche.

**Publications de référence**

**1 Alizon S.** C'est grave Dr Darwin ? L'évolution, les microbes et nous. 2016 Le Seuil, Paris.

**2 Grenfell B. T., Pybus O G, et al.** Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 2004; 303: 327-332.

**3 Volz EM,** Koelle K, Bedford T. Viral phylodynamics. *PLoS Comput Biol* 2013; 9

**4 Pybus OG,** Charleston M A, et al. The epidemic behavior of the hepatitis C virus. *Science* 2001; 292:2323-2325.

**5 Volz EM,** Kosakovsky Pond S L, et al. Phylodynamics of infectious disease epidemics. *Genetics* 2009; 183: 1421-1430.

**6 Stadler T,** Kühnert D, Bonhoeffer S, Drummond A J. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci USA* 2013; 110: 228-233.

**7 Lemey P,** Rambaut A, Drummond A J, Suchard M A. Bayesian phylogeography finds its roots. *PLoS Comput Biol* 2009; 5

**8 Rasmussen DA,** Boni M F, Koelle K. Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. *Mol Biol Evol* 2014;31 :258-271.

**Revue de la littérature**

**Frost S, Pybus O,** et al. Eight challenges in phylodynamic inference. *Epidemics* 2015;10:88-82.

**Du Plessis L, Stadler T.** Getting to the root of epidemic spread with phylodynamic analysis of genomic data. *Trends Microbiol* 2015;23(7):383-386.

**Autres publications identifiées**

**Villabona-Arenas C J, de Oliveira J L,** et al. Epidemiological dynamics of an urban Dengue 4 outbreak in São Paulo, Brazil. *Peer J.* 2016;4

*Cet article relate une approche phylodynamique visant à estimer le schéma de dispersion spatio-temporel du virus de*

*la dengue dans la ville de São Paulo, pendant l'épidémie de 2013. Les auteurs ont également estimé le taux de reproduction au cours du temps (R) grâce aux données d'incidence.*

**Kühnert D, Stadler T, Vaughan TG, Drummond AJ.** Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Mol Bio Evol* 2016; sous presses.

*Cet article présente une nouvelle méthode pour incorporer des processus de migration entre localités lors de l'analyse phylodynamique. Les auteurs appliquent leur nouvelle méthode au cas du virus influenza H3N2 circulant dans plusieurs pays et à celui du Virus de l'Immunodéficience Humaine circulant dans plusieurs groupes à risque.*

**Joy JB, McCloske RM,** et al. The spread of hepatitis C virus genotype 1a in North America: a retrospective phylogenetic study. *Lancet Inf Dis* 2016; sous presses

*L'origine de l'épidémie du virus de l'hépatite C aux USA demeure controversée. Cette étude apporte de nouvelles réponses à partir de données de séquences existantes en utilisant une approche phylodynamique pour dater le début de l'épidémie et retracer la dynamique du nombre d'infections.*

**Arafa A, El-Masry I,** et al. Phylodynamics of avian influenza clade 2.2.1 H5N1 viruses in Egypt. *Virology Journal* 2016; 13:49.

*Il s'agit d'une application récente de la phylodynamique à un jeu de données de 8 ans portant sur le virus de la grippe aviaire H5N1 en Egypte.*

**Bourhy H, Nakouné E,** et al. Revealing the micro-scale signature of endemic zoonotic disease transmission in an african urban setting. *PLoS path* 2016; 12(4):

*Cet article applique la phylodynamique à un jeu de données de 20 ans portant sur le virus de la rage chez les chiens domestiques d'Afrique centrale.*

#### Liens d'intérêts :

Les auteurs déclarent n'avoir aucun lien d'intérêt