Phylodynamique du COVID-19 en France : variations temporelles de la croissance épidémique

Gonché Danesh, Baptiste Elie, Samuel Alizon

Préambule

Phylodynamique

Données utilisées

Analyse via Beast

Discussion et limites

Logiciels utilisés

Sources et remerciements

Phylodynamique du COVID-19 en France : variations temporelles de la croissance épidémique

Groupe de modélisation de l'équipe ETE (https://www.mivegec.ird.fr/fr/contact/160-francais/equipes/1205-ete) (Laboratoire MIVEGEC (http://mivegec.ird.fr), CNRS, IRD, Université de Montpellier)

9 avril 2020 (mis à jour le 24 avril)

Préambule

Ce rapport a été produit à des fins académiques et ne constitue pas un support de prise de décision. De plus, les calculs réalisés ici sont faits avec des hypothèses volontairement simplifiées. Ainsi, les moyennes nationales ne reflètent pas nécessairement des situations locales.

L'ensemble de nos rapports sur la pandémie de COVID-19 est disponible sur cette page (http://covid-ete.ouvaton.org).

En matière de santé publique et pour toute question, nous recommandons de consulter et suivre les instructions officielles disponibles sur https://www.gouvernement.fr/info-coronavirus (https://www.gouvernement.fr/info-coronavirus)

L'Organisation Mondiale de la Santé (OMS) dispose aussi d'un site très complet https://www.who.int/fr/emergencies/diseases/novel-coronavirus-2019 (https://www.who.int/fr/emergencies/diseases/novel-coronavirus-2019)

Nous remercions les patients, personnels de santé et les laboratoires de virologie de France qui ont permis de générer les données de séquences de SRAS-Cov-2 à la base de ce travail. Pour plus de détails, voir la liste des données utilisées.

Phylodynamique

Ce rapport consiste à analyser des génomes viraux issus de patient·e·s infecté·e·s par le COVID-19 en France à l'aide d'outil de phylodynamique.

Pour plus de détails sur ces approches, on pourra se référer à notre précédent rapport sur la phylodynamique (Rapport4_Phylodynamique.html).

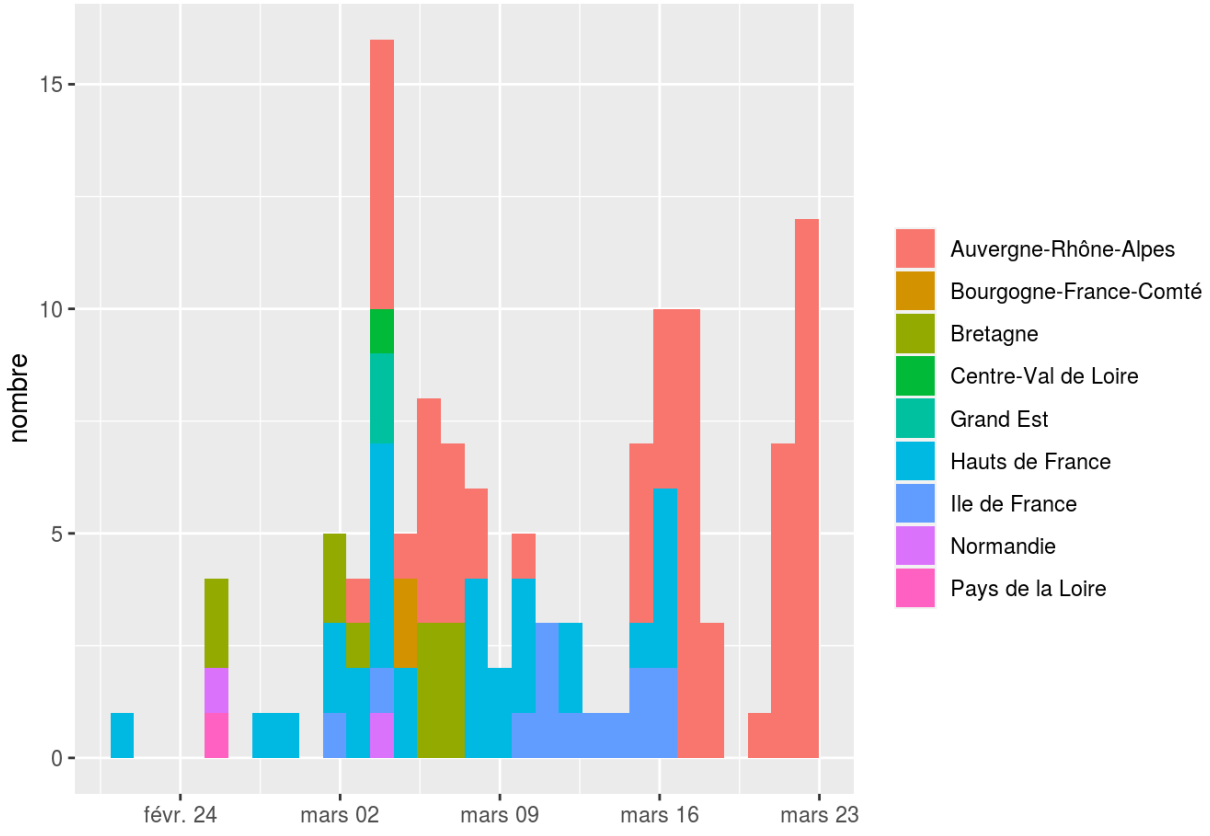
D'autres analyses ont déjà été réalisées au niveau mondial, qui sont disponibles sur le site virological.org (http://virological.org/). Le site nextstrain.org (https://nextstrain.org/) permet lui de suivre la propagation géographique de plusieurs épidémies virales quasiment en temps réel à l'aide des données de séquences disponibles.

Ce rapport se focalise sur le clade principal de l'épidémie en France en utilisant les données de SRAS-Cov-2 disponibles sur la base de données GISAID (https://www.gisaid.org/) au 3 avril 2020, soit `123` séquences pour la France. Ces séquences ne peuvent pas être publiées dans cette analyse car elles sont pour la plupart sous embargo.

Données utilisées

Comme indiqué dans notre Rapport 4 (Rapport4_Phylodynamique.html), la grande majorité des séquences provenant de France se regroupe dans le clade dénommé A2a par nextstrain et visible ici (https://nextstrain.org/ncov?f_country=France&label=clade:A2a).

Le graphique suivant représente les séquences utilisées, classées en fonction de la date à laquelle le prélèvement a été effectué et de la région de prélèvement.



Ces séquences ont été alignées et nettoyées à l'aide de la pipeline Augur développée par nextstrain (https://nextstrain.org/docs/bioinformatics/introduction-to-augur).

Nous avons analysé le jeu de données avec le logiciel RDP (https://web.cbio.uct.ac.za/~darren/rdp.html), qui n'a détecté aucun événement de recombinaison génétique.

Ce jeu de données, qui porte sur les séquences du plus gros cluster est divisé en 3 sous-ensembles :

1. 59 séquences qui étaient disponibles au 25 mars et dont la plus récente datait du 11 mars (jeu de donnée **France59**),
2. 42 séquence supplémentaires disponibles au 31 mars et dont la plus récente datait du 18 mars (jeu de donnée **France101**),
3. 21 séquences supplémentaires disponibles au 3 avril et dont la plus récente datait du 22 mars (jeu de données **France122**).

Analyse via Beast

Paramétrisation

Modèles d'évolution

Comme indiqué dans le Rapport 4 (Rapport4_Phylodynamique.html), afin de réaliser une analyse de phylodynamie il faut faire plusieurs hypothèses. Deux d'entre elles portent sur le modèle d'évolution de l'ADN (comment se font les mutations d'un nucléotide à un autre) et l'autre porte sur le taux de substitution (c'est-à-dire la vitesse à laquelle les mutations apparaissent et se fixent dans les séquences).

Afin de déterminer le modèle d'évolution le plus approprié étant donné notre alignement de séquences, nous avons utilisé le logiciel SMS (http://www.atgc-montpellier.fr/SMS/), qui identifie le modèle GTR (https://en.wikipedia.org/wiki/Substitution_model) comme le plus approprié selon le critère d'Akaike (AIC).

Concernant le taux de substitution (ou vitesse d'évolution) les deux options sont soit de le fixer à une valeur donnée, soit de tenter de l'estimer à partir des séquences datées. Avant de se lancer dans cette seconde approche, il faut cependant vérifier qu'il y a assez de signal dans les données. Ceci a été effectué à l'aide du logiciel TempEst (http://tree.bio.ed.ac.uk/software/tempest/) sur une phylogénie réalisée grâce au logiciel PhyML (http://www.atgc-montpellier.fr/PhyML/). Les résultats montrent les limites de ce jeu de données réduit au plus gros clade français.

En effet, comme indiqué dans le Rapport 4 (Rapport4_Phylodynamique.html), le coefficient de régression de la régression linéaire entre la distance à la racine de la phylogénie et la date d'échantillonnage, indique le taux de substitution. Ici, sa valeur de l'ordre de $7 \cdot 10^{-4}$ an^{-1} est cohérente avec les résultats obtenus par Andrew Rambaut sur l'ensemble de la phylogénie le 6 mars 2020 (http://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356). Toutefois, le coefficient de détermination, qui indique le pourcentage de la variance expliquée par la régression, est seulement de $R^2 = 6\%$, ce qui est peu (avec l'ensemble de la phylogénie il est de plus de 50 %).

En résumé, le signal phylogénétique dans les données est cohérent avec les estimations existantes. Toutefois, du fait du peu de séquences anciennes, il semble préférable de fixer le taux de substitution dans les analyses. Nous avons donc utilisé trois paramétrisations pour l'*horloge moléculaire* et donc la vitesse d'évolution : soit le taux de substitution a été fixé à $8,8 \cdot 10^{-4}$ substituions par génome par an (notre référence) ou à $13,2 \cdot 10^{-4}$ subst/position/an, soit il a été estimé (à partir d'une distribution a priori uniforme suivant une hypothèse d'horloge moléculaire stricte).

Modèles populationnels

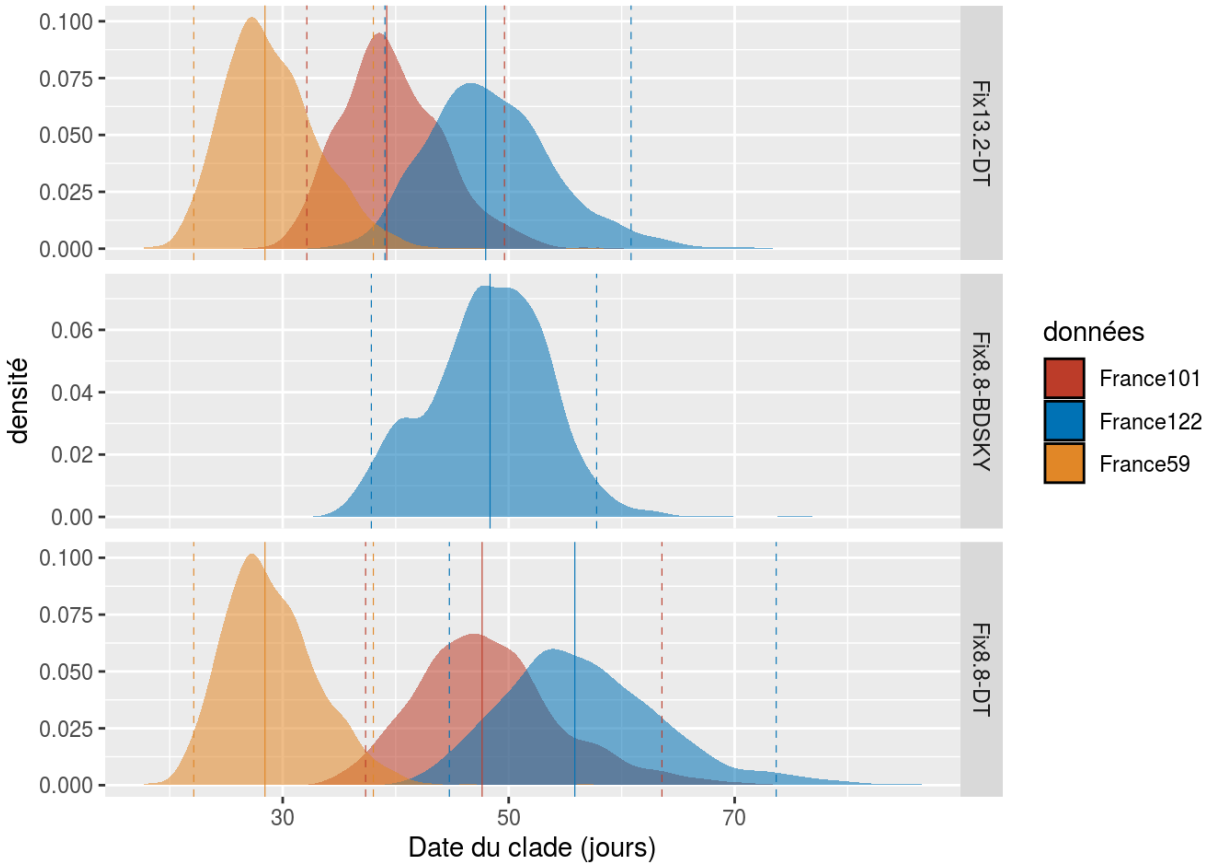
Nous avons réalisé les analyses phylodynamique à partir de notre alignement de séquences datées à l'aide des logiciels Beast (https://beast.community/) et Beast2 (http://www.beast2.org/). Sans entrer dans les détails, ces logiciels utilisent des approches bayésiennes afin d'intégrer des paramètres liés à un modèle d'évolution moléculaire (qui nous serviront à dater les événements) et à un modèle démographique (qui nous serviront à retracer la propagation). Nous avons analysé plusieurs modèles et présentons les résultats de deux d'entre eux :

1. Un modèle basé sur un coalescent exponentiel (noté **DT** pour *doubling time*), qui fait l'hypothèse que la population croit de manière exponentielle. À partir de ce modèle, nous pouvons notamment estimer le temps de doublement de l'épidémie, c'est-à-dire le nombre de jour qu'il faut attendre pour que le nombre de personnes infectées soit doublé. La limite de ce modèle est qu'il devient inadéquat si l'épidémie décroît, ce qui ne semble pas être le cas.
2. Un modèle basé sur un processus de naissance et de mort (noté **BDSKY**, pour *birth-death skyline plot*), qui est plus proche d'un modèle épidémiologique et où chaque nouvelle infection correspond à une "naissance" et chaque fin d'infection correspond à une "mort". À partir de ce modèle nous pouvons estimer à la fois le nombre de reproduction de base temporel (noté $\mathcal{R}(t)$ et décrit dans nos précédents rapports (Rapport5_R.html)) et la durée de contagiosité. En pratique, le temps est divisé en plusieurs périodes afin de détecter des variations de $\mathcal{R}(t)$. Une des limites de ce modèle est qu'il peut être sensible à l'échantillonnage et qu'il nécessite d'estimer de nombreux paramètres.

Datation

Grâce à l'inférence phylodynamique, on peut dater l'ancêtre commun à toutes les séquences de notre jeu de données. Biologiquement, cela peut-être interprété comme la date de l'infection qui a conduit à toutes ces infections détectées en France. Attention, il se peut tout à fait que cet ancêtre commun corresponde à une infection hors de France (par exemple en Chine), surtout s'il y a eu des introduction multiples dans le clade étudié.

Les datations proviennent d'un modèle de coalescent exponentiel (**DT**) réalisé sur les trois sous-jeux de données de tailles et de dates différentes (**Fra59**, **Fra101** et **Fra122**), ainsi que du modèle **BDSKY** réalisé uniquement sur le jeu de donnée le plus complet (**Fra122**). Dans tous les cas l'horloge moléculaire est fixée à une valeur moyenne ($8,8 \cdot 10^{-4}$) ou élevée ($13,2 \cdot 10^{-4}$).



On voit que les médianes (traits verticaux) sont proches quels que soient les jeux de données utilisés (de couleur différentes) et les modèles d'évolution (panneaux différents). La forme exacte des distributions varie, mais les intervalles de confiance sont très similaires.

Le tableau ci-dessous indique les dates pour une valeur de vitesse d'évolution moyenne ($8,8 \cdot 10^{-4}$ subst/position/an).

Modèle	Nombre de séquences	Date la plus récente	borne inf. (95 %)	médiane commun	borne sup. (95 %)
Fix8.8-DT	59	11 mars 2020	01 févr. 2020	11 févr. 2020	17 févr. 2020
Fix8.8-DT	101	18 mars 2020	14 janv. 2020	30 janv. 2020	09 févr. 2020
Fix8.8-DT	122	22 mars 2020	08 janv. 2020	26 janv. 2020	06 févr. 2020
Fix8.8-BDSKY	122	22 mars 2020	24 janv. 2020	02 févr. 2020	13 févr. 2020

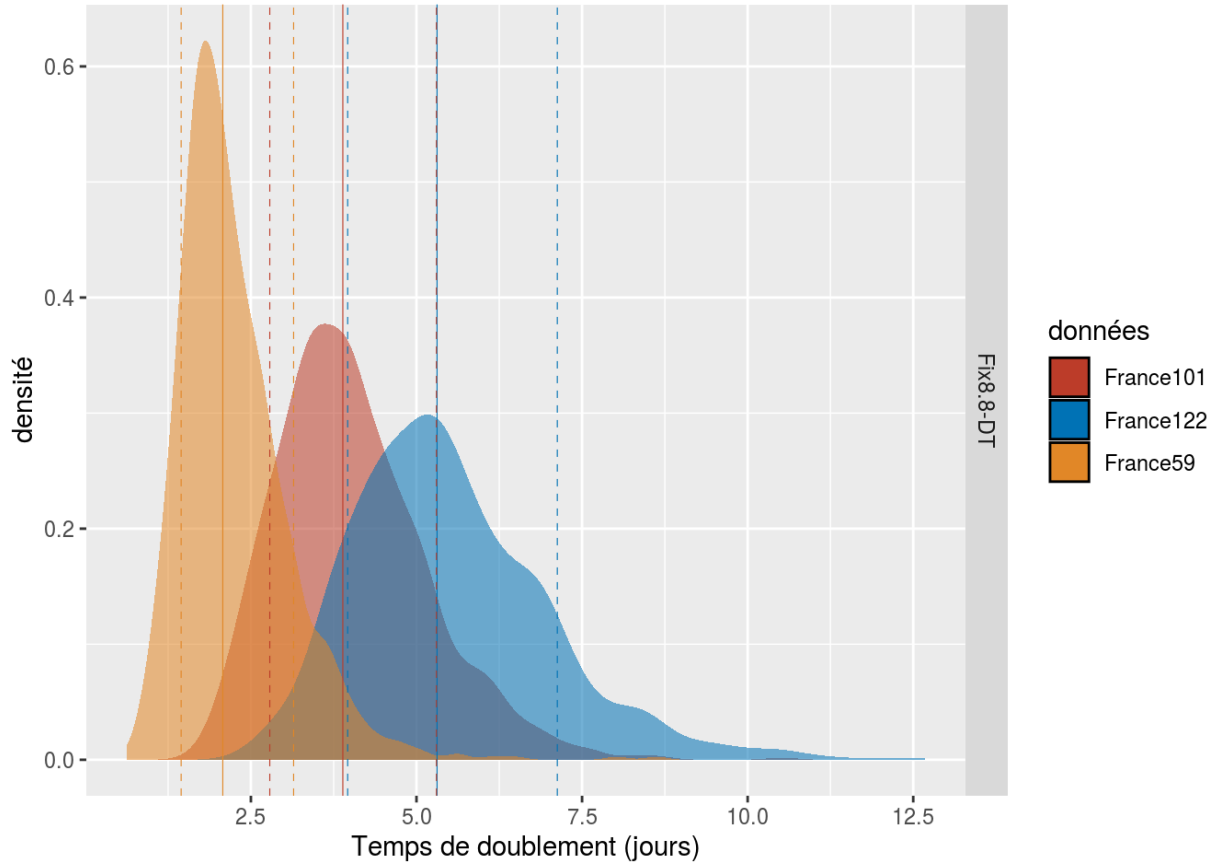
On voit qu'avec le jeu de données le plus complet on date l'ancêtre commun à ce principal clade en France entre la mi-janvier et la mi-février. Ce grand intervalle s'explique d'une part par le peu de séquences "anciennes" (la première récoltée dans ce clade date du 21 février) ainsi que par le fait que nous moyennons ici toutes les régions de France, qui ont pu connaître des introductions indépendantes provenant de différents pays.

Ces dates sont cohérentes toutefois avec celles obtenues par Andrew Rambaut (http://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356) quant au début de l'épidémie en Chine qui est daté au 17 novembre 2019 avec un intervalle de confiance entre le 27 août et le 19 décembre 2020.

Temps de doublement

L'étude de la phylodynamique nous permet d'estimer des valeurs de paramètres de dynamique de population. Pour un modèle de coalescent avec croissance exponentielle, le paramètre clé est le temps de doublement, qui correspond au nombre de jours pour que l'épidémie double en taille. Ce paramètre sert à calculer le R_0 comme nous l'expliquons dans notre Rapport 5 (Rapport5_R.html).

Nous avons mesuré ce taux de croissance pour nos trois sous-ensembles. Comme le jeu de données **France122** inclut des séquences plus récentes que **France101**, qui lui-même inclut des séquences plus récentes que **France59**, notre hypothèse est que nous pouvons détecter des variations du temps de doublement au cours de l'épidémie.



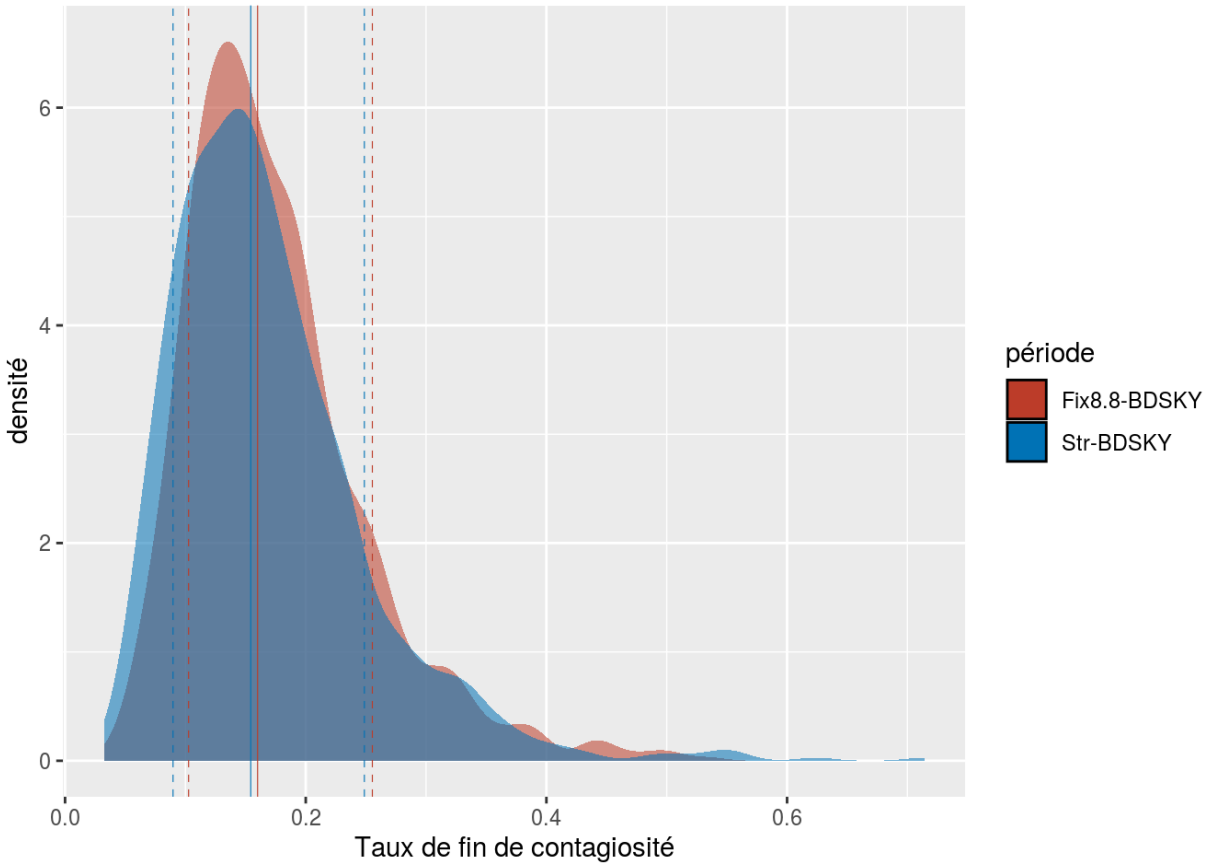
Comme on le voit sur la figure, en rajoutant des données de séquence plus récentes à notre jeu de données, on observe une augmentation du temps de doublement de l'épidémie. Initialement, avec les 59 premières séquences (qui vont du 21 février au 11 mars), le temps de doublement était très rapide de l'ordre de 2,5 jours. En rajoutant des séquences échantillonnées entre le 12 et le 18 mars, ce temps monte à 4 jours. Enfin, en rajoutant les séquences échantillonnées entre le 19 et le 22 mars, on passe au dessus de 5 jours.

En comparaison, les inférences phylodynamiques réalisées à partir des données de Chine (avec 86 génomes, Andrew Rambaut (http://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356)) trouvait un temps de doublement médian d'environ 7 jours avec un intervalle de confiance entre 4.7 et 16.3 jours). Ces faibles vitesse de croissance de l'épidémie par rapport aux notres s'expliquent par le fait que nous nous sommes concentrés sur un clade de l'épidémie en expansion rapide et avons négligé les clades moins gros.

Durée de contagiosité

Le modèle BDSKY permet d'aller plus loin de deux manières. D'une part, il permet d'accéder à une durée de contagiosité. D'autre part, il permet d'estimer le nombre de reproduction de l'épidémie (c'est-à-dire le nombre d'infections secondaires engendrées par une personne infectée). Le modèle de coalescent exponentiel décrit ci-dessus n'accédait lui qu'au quotient entre ces deux termes.

À noter que le modèle BDSKY nécessite d'estimer plus de valeurs de paramètres. De plus il faut aussi estimer le taux d'échantillonnage après le 21 février (date à laquelle a été échantillonnée la séquence la plus ancienne) alors que le modèle de coalescent suppose que cet échantillonnage est négligeable.



Les taux de fin de contagiosité sont très similaires selon que la vitesse d'évolution est fixée à une valeur donnée ($8,8 \cdot 10^{-4}$ subst/position/an) ou estimée.

À partir de ce taux, on peut obtenir la durée de contagiosité, sachant qu'il faut aussi rajouter le taux d'échantillonnage car *a priori* les patient·e·s dont les infections ont été séquencées ont peu transmis après cette détection. Le taux d'échantillonnage est estimé à `0.057` jours^{-1} avec un intervalle de confiance à 95 % allant entre `0.011` et `0.292` .

La distribution des durées de contagiosité est obtenue en prenant l'inverse de la somme du taux d'échantillonnage et du taux de fin de contagiosité (l'inverse d'un taux est une durée). La médiane de cette distribution est `4.17` jours et 95 % de ses valeurs sont comprises entre `2.22` et `5.94` jours.

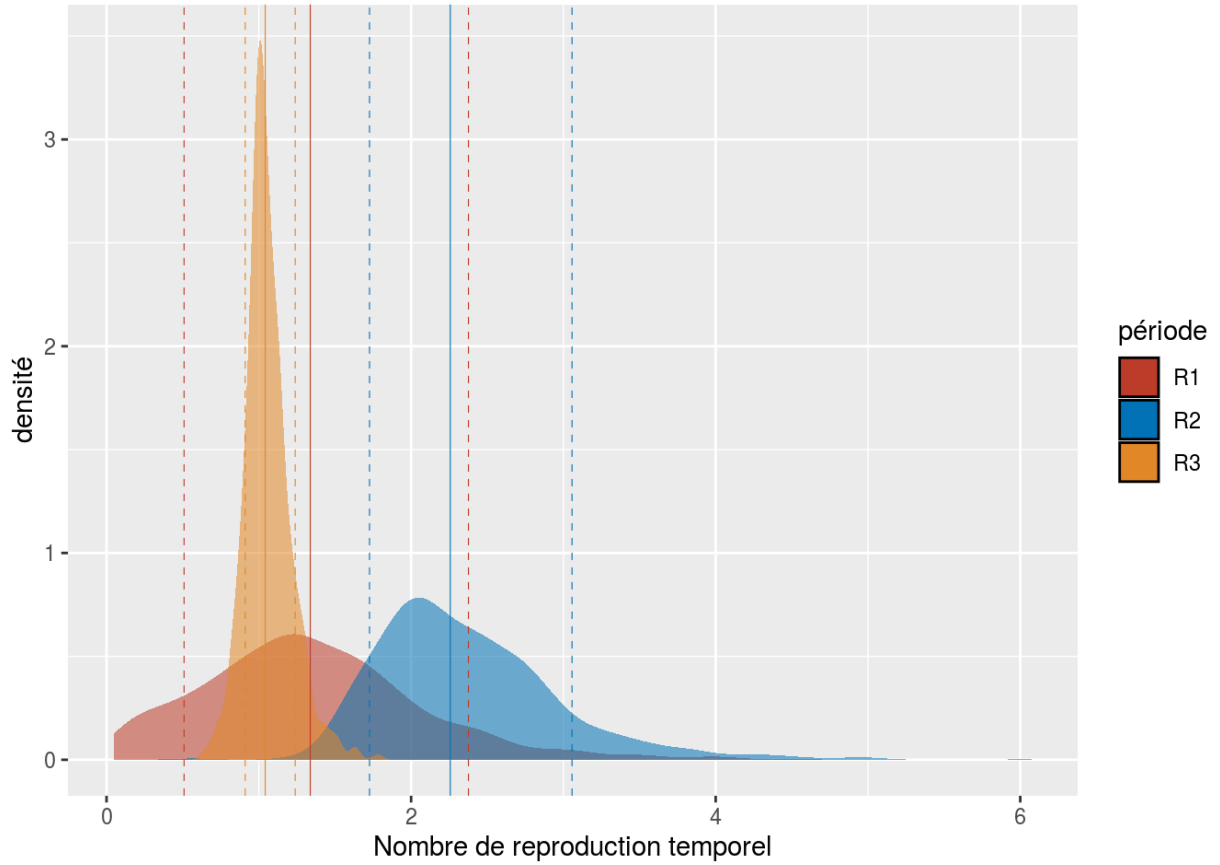
Les valeurs que nous estimons à l'aide des séquences datées sont particulièrement cohérentes avec celles inférées à partir des études épidémiologiques de suivi de contact. Ces dernières mesurent des intervalles sériels, qui servent à approximer cette durée de période de contagiosité dans le calcul du $\mathcal{R}_0$. Ainsi, Nishiura *et alii* (2020, *Int J Infect Dis*) trouvent une distribution dont la médiane est 4 jours et l'intervalle de confiance à 95 % entre 3,1 et 4,9 jours.

Nombre de reproduction

En parlant de nombre de reproduction de base, on peut estimer sa valeur à un moment donné et calculer le nombre de reproduction efficace noté $\mathcal{R}(t)$. Ici, nous avons divisé le temps en 3 intervalles afin de détecter des variations de $\mathcal{R}(t)$:

- $\mathcal{R}_1$ estime le nombre de reproduction entre le 2 et le 18 février,

- $\mathcal{R}_2$ estime le nombre de reproduction entre le 18 février et le 6 mars,
- $\mathcal{R}_3$ estime le nombre de reproduction entre le 6 et le 22 mars.



Ces résultats sont très cohérents avec ceux obtenus pour le temps de doublement, même si les périodes temporelles sont différentes. Concernant la période avant le 18 février, l'estimation est la moins précise avec des valeurs de $\mathcal{R}_1$ comprises à 95 % entre `0.15` et `3.05`. Ceci est cohérent avec le fait que la séquences la plus ancienne date du 21 février alors que la racine de l'arbre est estimée à début février. Sur la deuxième période temporelle, on détecte une croissance rapide puisque les valeurs de $\mathcal{R}_2$ comprises à 95 % entre `1.54` et `3.86`. Enfin, la période la plus récente détecte un ralentissement de l'épidémie avec un $\mathcal{R}_3$ compris à 95 % entre `0.79` et `1.43`.

Discussion et limites

L'analyse de séquences virales dont la date de prélèvement est connue permet de réaliser des phylogénies d'infections et d'estimer des valeurs de paramètres épidémiologiques. Nous avons réalisé cette analyse en nous basant sur les séquences représentant le plus gros clade au sein de la phylogénie de l'ensemble des séquences d SRAS-Cov-2 disponibles. Ce clade peut être visualisé sur la représentation de nextstrain (https://nextstrain.org/ncov?f_country=France&label=clade:A2a).

Avant de résumer les résultats, nous préférons pointer plusieurs limites de notre analyse :

- le clade français que nous avons analysé est en fait un clade européen (voire international puisqu'on trouve aussi des séquences US) : bien que l'on distingue dans ce clade deux gros "cluster" avec uniquement des séquences françaises, il est possible que les variations de vitesse de croissance que nous détectons soient plus dues à des politiques de contrôle européennes que françaises ;

- les séquences collectées du 17 au 22 mars (les plus récentes) proviennent toutes de la même région (Auvergne-Rhône-Alpes) comme on le voit sur notre première figure : là encore, les variations détectées récemment pourraient être dues à l'épidémie dans cette région en particulier ;
- l'horloge moléculaire a dû être fixée dans cette analyse car nous ne disposons pas d'un échantillonnage suffisant au cours du mois de février en France.

Malgré ces limites, nos résultats suggèrent un ralentissement de l'épidémie en France. Ainsi, entre le 11 et le 22 mars, le temps de doublement de l'épidémie estimé par un modèle de *coalescent de croissance exponentielle* a plus que doublé, passant d'une médiane inférieure à 2,5 jours à une médiane supérieure à 5 jours. Ce ralentissement est aussi détecté à l'aide d'un modèle de naissance et de mort via le nombre de reproduction temporel $\mathcal{R}(t)$: à partir du 7 mars, on passe d'une valeur médiane de `2.27` à une valeur médiane de `1.04`. Ceci est cohérent avec la mise en place d'un confinement en France à compter du 16 mars. Ces variations et même ces ordres de grandeurs sont cohérents avec nos estimations réalisées à partir des séries temporelles d'incidence de nouvelles hospitalisations et de décès dans notre Rapport 5 (Rapport5_R.html).

Enfin, le modèle BDSKY permet aussi d'obtenir une estimation de la durée de contagiosité, qui est essentielle dans le calcul du $\mathcal{R}_0$. La valeur médiane obtenue est plutôt élevée (`6.25` jours) mais cohérente avec certaines estimation d'intervalle sériel faite en Asie. À ce jour, il n'existe d'ailleurs pas d'estimation de cette durée en France.

En augmentant le nombre de séquences génomiques de SARS-CoV-2 issues de l'épidémie française (et le nombre de personnes travaillant sur le sujet), en particulier des séquences prélevée en début d'épidémie, il serait possible de :

- mieux dater l'arrivée de l'épidémie en France,
- mieux comprendre la propagation entre les différentes régions de France,
- estimer le nombre d'introductions du virus dans le pays,
- détecter de potentielles mutations adaptatives qui modifieraient la propagation du virus.

Pour aller plus loin sur la phylodynamique…

- Notre Rapport 4 (Rapport4_Phylodynamique.html), qui décrit l'épidémie en France au 18 mars.
- https://nextstrain.org/ncov (https://nextstrain.org/ncov) : le nec plus ultra de la visualisation phylodynamique au niveau mondial
- Volz *et alii* (2020) (https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-phylogenetics-15-02-2020.pdf) : le rapport du groupe de modélisation partenaire de l'OMS (en anglais).
- http://virological.org/ (http://virological.org/) : site qui regroupe des analyses et des discussions en phylogénétique et phylodynamique avec en particulier l'analyse d'Andrew Rambaut (http://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356) et celle du groupe de Tanja Stadler (http://virological.org/t/update-2-evolutionary-epidemiological-analysis-of-128-genomes/423) très axée sur la phylodynamique.

Logiciels utilisés

- L'alignement des séquences a été obtenu par la pipeline Augur développée par nextstrain (https://nextstrain.org/docs/bioinformatics/introduction-to-augur).

- Le choix du modèle génétique de substitution a été réalisé à l'aide du logiciel SMS (http://www.atgc-montpellier.fr/sms/).
- La phylogénie en maximum de vraisemblance a été inférée à l'aide du logiciel PhyML (http://www.atgc-montpellier.fr/PhyML/).
- La recherche de signal temporel d'évolution moléculaire a été réalisée via le logiciel TempEst (http://tree.bio.ed.ac.uk/software/tempest/).
- Les analyses de phylodynamique ont été réalisée grâce à la version 1.8.3 du logiciel Beast (https://beast.community/) et à la version 2.3 du logiciel Beast2 (http://www.beast2.org/).
- Des manipulations additionnelles ont été effectuées en R (http://cran.r-project.org/) grâce au package ape (http://ape-package.ird.fr/).

Sources et remerciements

- Les données de séquences génétiques ont été fournies par plusieurs laboratoires internationaux et mis à disposition au travers de GISAID - Global Initiative on Sharing All Influenza Data (https://www.gisaid.org/). La liste des séquences utilisées et les laboratoires les ayant obtenues sont dans le tableau ci-dessous.

accession <fctr>	**date** <fctr>	**city** <fctr>
EPI_ISL_418415	2020-03-15	ARA
EPI_ISL_418414	2020-03-15	ARA
EPI_ISL_418413	2020-03-15	ARA
EPI_ISL_418412	2020-03-15	ARA
EPI_ISL_418419	2020-03-16	ARA

1-5 of 210 rows | 1-3 of 6 columns Previous **1** 2 3 4 5 6 … 42 Next

- Nous remercions les patients, personnels de santé et les laboratoires de virologie de France qui ont permis de générer les données de séquences de SRAS-Cov-2 à la base de ce travail.
- Les auteurs remercient le calculateur à haute performance itrop (plateforme South Green (http://www.southgreen.fr)) de l'IRD de Montpellier pour la fourniture des ressource de calcul à haute performance, qui ont contribué aux résultats présentés dans ce travail (plus de détails sur bioinfo.ird.fr (https://bioinfo.ird.fr/)).
- Gonché Danesh est financée par une bourse doctorale de la Fondation pour la Recherche Médicale (http://www.frm.org).

- L'équipe de modélisation de l'équipe ETE est composée de Samuel Alizon, Thomas Bénéteau, Marc Choisy, Gonché Danesh, Ramsès Djidjou-Demasse, Baptiste Elie, Yannis Michalakis, Bastien Reyné, Quentin Richard, Christian Selinger, Mircea T. Sofonea.
- Contribution à ce travail :
    - conception du travail : toute l'équipe
    - réalisation des analyses : GD, BE et SA
    - rédaction du rapport : SA

    - contacts : samuel.alizon@cnrs.fr (mailto:samuel.alizon@cnrs.fr), gonche.danesh@ird.fr (mailto:gonche.danesh@ird.fr), baptiste.elie@ird.fr (mailto:baptiste.elie@ird.fr)
    - approbation : toute l’équipe
    - Remerciements à Andrew Rambaut pour sa clarification de points liés à l’interprétation des résultats.
-