CNR Virus des infections respiratoires

Cahier des charges pour l'ouverture du séquençage aux laboratoires de biologie médicale (LBM) dans le cadre de la stratégie nationale de surveillance génomique du SARS-CoV-2 (projet EMERGEN).

Version du 10/05/2021.

La stratégie nationale de surveillance génomique du SARS-CoV-2 (projet EMERGEN) est coordonnée par Santé publique France et l'ANRS|Maladies Infectieuses Emergentes (MIE). Elle associe de nombreux partenaires et institutions, dont le Centre National de Référence (CNR) Virus des infections respiratoires et l'Institut Français de Bioinformatique (IFB).

Elle a pour objectif de décrire et suivre la circulation des variants déjà connus, mais aussi et surtout de détecter, identifier, évaluer et suivre dans les meilleurs délais la circulation de nouveaux variants préoccupants (c'est-à-dire ceux ayant des conséquences fonctionnelles, en termes de transmissibilité ou de pathogénicité par exemple). Elle s'appuie pour cela sur une liste de cibles prioritaires (indications du séquençage) permettant d'explorer différentes populations (en ville, en établissements médico-sociaux, en établissements de santé) et d'obtenir le matériel biologique et les métadonnées nécessaires aux activités de séquençage et de leur interprétation.

Les prérequis en matière de conformité des techniques et des remontées des résultats de séquençage dans la base de données nationale du projet ont été définis par Santé publique France, l'ANRS|MIE, l'Institut Français de Bioinformatique et le CNR Virus des infections respiratoires. Santé publique France et l'ANRS|MIE ont par ailleurs défini des indications prioritaires pour le séquençage du SARS-CoV-2, parmi lesquelles figurent le séquençage à visée de surveillance (sur la base d'un échantillonnage aléatoire). Ces indications ont fait l'objet d'un MINSANTE n°48-2021 du 30/03/2021.

L'ensemble des données de séquençage du SARS-CoV-2 produites dans le cadre de cette stratégie ont vocation à alimenter les missions de surveillance, d'investigation et d'évaluation des risques de Santé publique France et les travaux de recherche coordonnés par l'ANRS|MIE. Santé publique France et l'ANRS|MIE définissent les conditions d'accès des données ainsi recueillies.

Le présent cahier des charges définit les attendus des prestations de séquençage fournies par les laboratoires de biologie médicale (LBM) quelles que soient leurs indications :

- Séquençage à des fins de surveillance (sélection aléatoire) pour les LBM sélectionnés dans le cadre de l'appel à manifestation d'intérêt (AMI) conduit par Santé publique France ;
- Séquençage à visée interventionnelle pour les LBM sollicités sous couvert des ARS ;
- Séquençage suite à dépistage positif aux frontières selon les termes de l'article 28-1 de l'arrêté du 10 juillet 2020.

Ces activités de séquençage ouvrent droit pour les LBM sélectionnés à rémunération sur la base de la Nomenclature des Actes de Biologie Médicale (NABM). Le respect de ce cahier des charges et des volumes de séquençage concernés par ces indications feront l'objet d'un reporting hebdomadaire à l'Assurance Maladie, sur la base des données effectivement remontées dans la base de données nationale du projet EMERGEN.

1. Cahier des charges

A. Qualité du séquençage génome entier par NGS

Le séquençage NGS, quelle que soit la technique utilisée (Illumina, Ion Torrent, MinION), est la méthode de référence, et doit couvrir la totalité du génome viral.

D'après l'évaluation de méthode de séquençage génome entier SARS-CoV-2 du CNR (Charre, Virus Evol, 2020), les critères pour valider une séquence sont les suivants :

- Couverture minimale pour appeler une base : >10X pour Illumina et Ion Torrent; > 20X pour MinION. En dessous de ce seuil, la base n'est pas déterminée (i.e. appelée N) ;
- Séquence validée si >99% de la séquence est déterminée (i.e. sans base ambiguë N) avec une profondeur de couverture moyenne de 1000x ;
- Pour les délétions : il est indispensable de différencier délétion et défaut de couverture. Une délétion est visible sur le fichier d'alignement (cf Charre, Virus Evol, 2020) ;
- Au moins un contrôle négatif qui a suivi l'ensemble du pipeline de production doit être séquencé pour chaque série d'extraction/d'amplification génique afin de déterminer la présence ou non de contamination.

Un variant est défini à la fois par son clade selon la nomenclature Nextclade et son lignage selon la nomenclature Pangolin en vigueur à la date du dépôt des données sur le serveur EMERGEN. En outre, la liste exhaustive des mutations génomiques et/ou protéiques pourra être requise, selon un format spécifié par le consortium (actuellement, le format d'export du logiciel Nextclade).

Les mutations (substitutions et délétions) sont définies par rapport à la séquence de référence Wuhan/Hu1/2019 (NCBI Nucleotide – NC_045512, GenBank – MN908947) (Wu et al. Nature. 2020 Mar;579(7798):265-269).

B. Qualité des métadonnées associées

Suite à un acte de séquençage, les données transmises doivent inclure les métadonnées suivantes, conformément aux procédures du projet EMERGEN :

- Phase 1 de l'infrastructure du projet (actuellement en vigueur) :

Un tableau de résultats de catégories de virus en format .xlsx respectant la structure et les indications d'un fichier-trame disponible sur le serveur EMERGEN-DB (https://emergen-db.france-bioinformatique.fr/).

La version actuelle de ce fichier contient les champs suivants :

- Numéro de prélèvement (propre au laboratoire) – Alphanumérique ;
- Indication du séquençage – cluster, enquête Flash, voyage étranger ou contact avec voyageur, sujet vacciné, réinfection, immunodéprimé, cas graves, enquêtes dépistage, échec traitement Ac, situation épidémiologique anormale, surveillance sentinelles ;
- Département de résidence – 3 chiffres ;
- Année de naissance – 4 chiffres ;
- Date de prélèvement – AAAA/MM/JJ ;
- Nom du laboratoire préleveur – Alphanumérique ;
- Code postal du laboratoire préleveur – 5 chiffres ;
- Nom du laboratoire séquenceur – Alphanumérique ;
- Date de réception du prélèvement – AAAA/MM/JJ ;
- Date de validation du résultat du séquençage – AAAA/MM/JJ ;
- Type de séquençage – NGS obligatoirement ;
- Longueur de séquençage – Complet obligatoirement
- Résultat de la séquence :
    - Clade selon la nomenclature Nextstrain comme SIDEP ou VAR_IND ou PREL_NC ;
    - Lignage selon la nomenclature Pangolin ;
    - Liste des mutations identifiées.

Les spécifications de ce fichier pourront éventuellement évoluer au fil du projet.

- <u>Phase 2 (à partir de mai 2021)</u> :
    - Les séquences brutes (format fastq) en ayant pris soin d'éliminer les éventuels fragments de génome humain ;
    - La séquence assemblée (consensus) ;
    - Une description formelle (scripts, workflows) des pipelines utilisés pour extraire les consensus et variants à partir des données brutes, suffisamment complète et détaillée pour assurer la reproductibilité des résultats ;
    - Les métadonnées nécessaires à la soumission des séquences à GISAID (pour les séquences consensus) et EBI-ENA (pour les séquences brutes) ;

- <u>Phase 3 (été 2021, sous réserve de l'obtention des autorisations CNIL et certification de l'entrepôt de données par l'ANS)</u> :
    - Nom, prénom, date de naissance, sexe et code postal de résidence du patient.

Les calendriers d'implémentation des phase 2 et 3 seront confirmés ultérieurement. Une transmission rétroactive des données additionnelles sera alors demandée.

C. Traitement et transmission des données

Santé publique France et l'Inserm/ANRS-MIE sont responsables du traitement des données à caractère personnel relatif à l'entrepôt EMERGEN-DB centralisant, à des fins de surveillance et de recherche, des données existantes issues des systèmes d'information des laboratoires et les données de séquences produites.

Les modalités de transmissions des données précédemment décrites sont définies par Santé publique France et l'Inserm/ANRS|MIE, en lien avec l'Institut Français de Bioinformatique (IFB).

Ces données doivent être transmises vers la base nationale maintenue dans le cadre du projet EMERGEN et hébergée par l'IFB :

- Selon les procédures de soumission définies pour le projet en matière d'ouverture de comptes, de formats de fichier, de téléversement des données, de validation de la conformité des données et métadonnées. Ces procédures sont disponibles sur le site de l'IFB (https://emergen-db.france-bioinformatique.fr/) ;
- Sous 7 jours ouvrés maximum après la date du prélèvement.

Les LBM habilités par Santé publique France suite à l'AMI pour le séquençage à des fins de surveillance, ainsi que les LBM habilités par les ARS pour le séquençage à visée interventionnelle, demeurent responsables des traitements de données relatifs à la collecte initiale des données transmises, ainsi que de la conservation et de la réutilisation des données et séquences pour des finalités autres que la transmission vers l'entrepôt de données EMERGEN.

D. Enregistrement dans la base de données GISAID

Dès que la phase 2 de l'infrastructure EMERGEN sera opérationnelle, la soumission à GISAID des résultats de séquençage produits par les LBM sera faite par l'IFB, qui soumettra ces séquences en leur nom.

E. Alerte

En attendant la mise en œuvre de la phase 3 de l'infrastructure du projet EMERGEN et son interconnexion avec la base SI-DEP, les résultats de séquençage accompagnés de l'identité du patient doivent être transmis sans délai par le LBM aux ARS et Cellules régionales de Santé publique France concernées et saisis dans SI-DEP (champ JOKER3).

En cas de détection de nouvelles mutations susceptibles d'avoir une signification fonctionnelle ou de nouvelles mutations récurrentes (en référence aux analyses de risque publiées sur le site de Santé publique France), il est également demandé au LBM, spontanément ou sur demande du CNR ou de Santé publique France, d'envoyer les prélèvements correspondants au CNR.

Ce besoin d'un envoi éventuel au CNR rend nécessaire pour les LBM qui réalisent des actes de séquençage de conserver les prélèvements concernés un temps suffisant dans leurs collections.