N°28146

Base de données innovantes en génétique humaine : SCNVBase, pour stocker les variations du génome humain

Kuan-Hua Artignan[1,2], Philippine Garret[1,2], Cyril Fournier[1,5], Christophe Philippe[1,2], Laurence Faivre[1,2,3], Christel Thauvin-Robinet[1,2,4], Yannis Duffourd[1,2]

1. UMR1231 Inserm Equipe GAD – Univ. Bourgogne-Franche Comté – Dijon – France, 2. Unité Fonctionnelle Innovation en Diagnostic génomique des maladies rares – FHU-TRANSLAD – CHU Dijon Bourgogne – France, 3. Centre de Référence maladies rares « Anomalies du développement et syndromes malformatifs » – centre de génétique – FHU-TRANSLAD – CHU Dijon Bourgogne – France 4. Centre de Référence maladies rares « déficiences intellectuelles de cause rare » – centre de génétique – FHU-TRANSLAD – CHU Dijon Bourgogne 5. UMR1231 Inserm Equipe SAPHIHR

INTRODUCTION

L'essor du séquençage à haut-débit améliore grandement le taux de diagnostic des maladies rares et la connaissance du génome humain. Mais ces nouvelles technologies génèrent des quantités importantes de données. Le stockage et l'interrogation de ces données hétérogènes par leurs formats reste une difficulté importante pour agréger les variations détectées afin de découvrir de nouveaux gènes impliqués en pathologie humaine. Il est donc nécessaire d'inventer de nouveaux moyens pour améliorer le stockage et l'analyse de ces données.

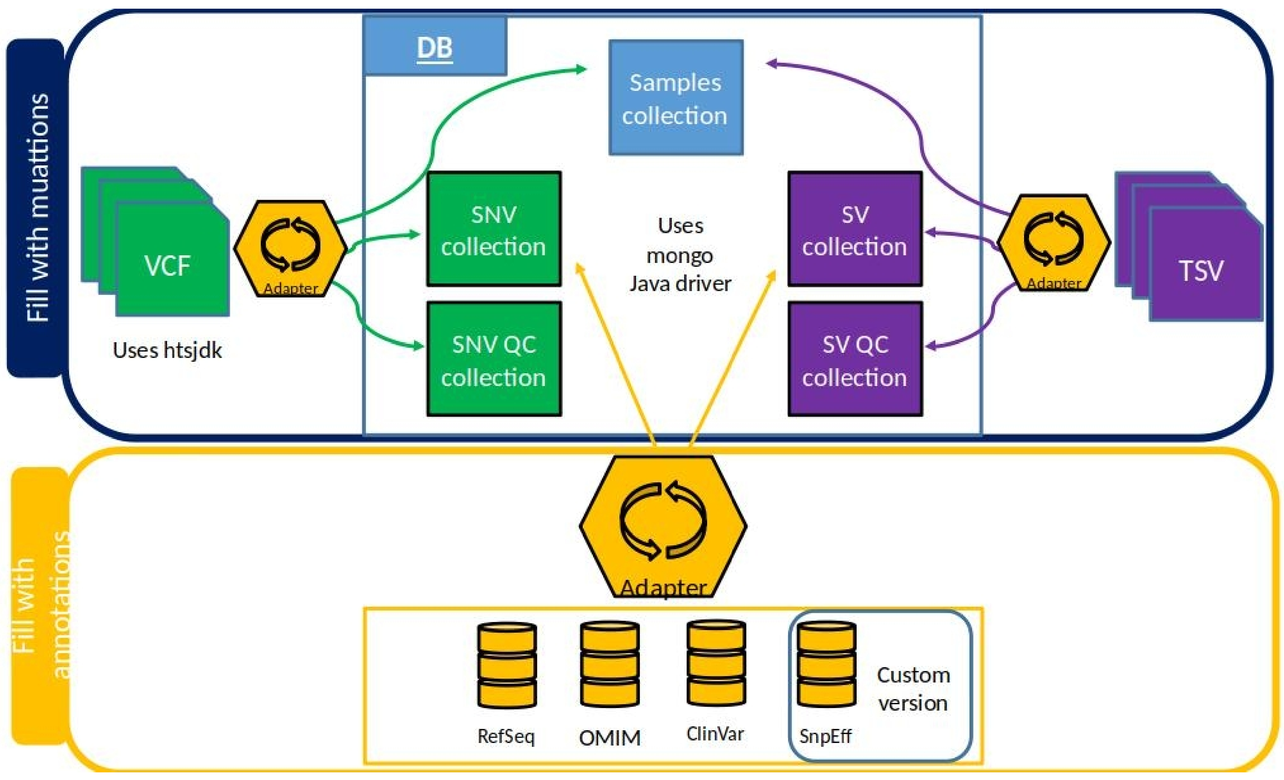

Figure 1 - Fonctionnalités de l'outil SCNVBase

METHODES

Une base de données innovante, SCNVBase, a été créée pour permettre le stockage et la recherche efficace dans ces données volumineuses. Elle est basée sur un système de gestion de base NoSQL, MongoDB. Ce système est donc bien adapté au stockage et à l'interrogation de grands volumes de données hétérogènes. Nous avons créé un outil en Java pour insérer les données concernant les variations de type CNV et SNV et les annoter rapidement avec des bases externes telles RefSeq ou OMIM via l'intégration de l'outil SnpEff. Une architecture 3-tiers a été utilisée. Le web service récupère les données de la base et les envoie à l'interface web. Le web service incluant la base MongoDB a été fait en Python, Flask. Une couche de sécurité est ajoutée par Flask JWT utilisant un système de jeton. L'accès aux données anonymisées est libre et sans authentification, permettant d'utiliser les informations de fréquence populationnelle. Relier les variations aux patients et à leurs phénotypes est possible après authentification de l'utilisateur et l'autorisation préalable par l'administrateur. Ce système permet un accès fiable et sécurisé aux données identifiantes.

RESULTATS

Le web service récupère les données dans la base et les envoie à l'interface web. Nous pouvons interroger les données en les visualisant sous forme de table ou de graphiques comme Lollipop et à la fois les champs à visualiser sont sélectionnables selon la volonté de l'utilisateur. La requête peut porter sur 1) le nom du gène, 2) la position et 3) la variation. La première version de l'application est déployée sur nos serveurs en test interne et sera bientôt mise à disposition des chercheurs[a]. Le modèle de la base de données est libre et accessible sur notre dépôt git[b]

a. https://scnvbase.gad-bioinfo.org
b. https://gitlab.gad-bioinfo.org/gad-public/scnvbase

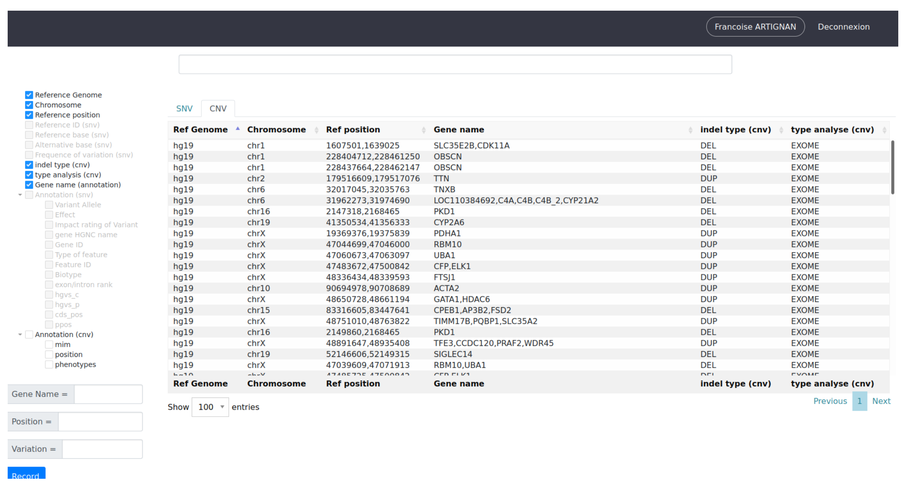

Figure 2 – Interface graphique pour la recherche interactive de variation

CONCLUSION

Le système affiche de très bonnes performances permettant l'interaction en temps réel avec les données. Les tests montrent un temps de réponse constant de quelques millisecondes pour un jeu de données augmentant incrémentalement.
De nouvelles fonctionnalités seront intégrées au fur et à mesure de l'utilisation. Ainsi, de nouveaux types de variation vont être supportés et intégrés à la base (Eléments mobiles, STR, etc.) et de nouvelles solutions d'interrogation seront mises à disposition. La base, ainsi que le modèle de la base restera libre, gratuite et accessible à tous grâce à l'utilisation d'une licence GPLv3.

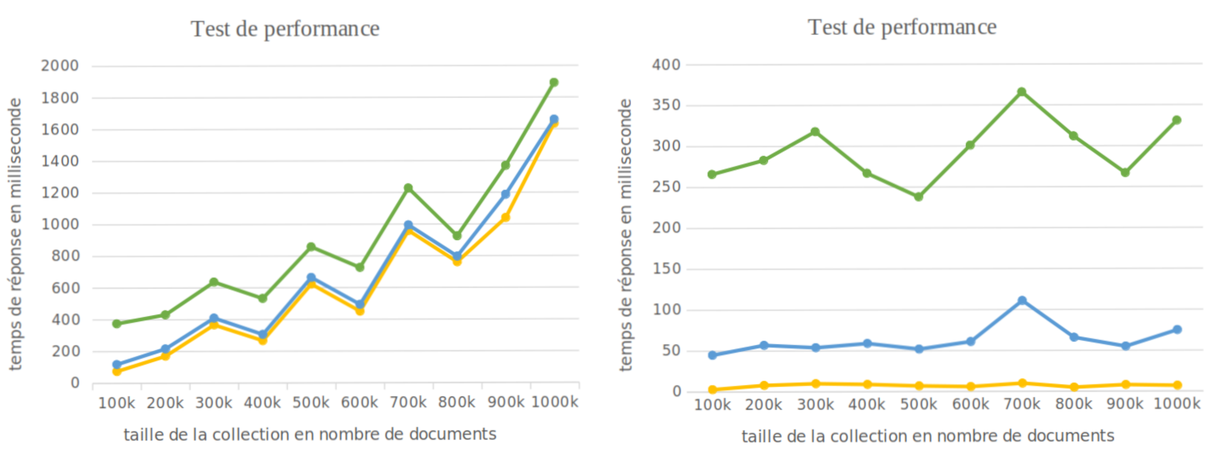

Figure 3 - Tests de performance dans « le pire des cas » (partie gauche) et dans « le meilleur des cas » (partie droite)