1. Objectifs du pipeline bioinformatique STARK

Le pipeline bioinformatique doit être capable de détecter les variations suivantes sur les régions cibles +/- 50 bases par rapport aux bornes définies lors du design d'un panel (fichier *bed*).

- SNV (single nucleotide variant) à l'état homozygote ou hétérozygote
- Petites indel (insertion ou délétion) de moins de 20 bases à l'état homozygote ou hétérozygote
- CNV (copy number variant ou variations du nombre de copies) à l'état homozygote ou hétérozygote

Le pipeline, nommé STARK, doit identifier les variations après alignement sur le génome hg19 et donner un statut (homozygote ou hétérozygote). La détection des variations en mosaïque et somatique est exclue de cette qualification.

Les termes techniques informatiques et bioinformatiques sont décrits au travers des recommandations émises par l'ANPGM (« Recommandation générale pour la gestion informatique des données et des analyses de séquençage à haut débit pour les laboratoires de diagnostic moléculaire de maladies génétiques »).

2. Développement du pipeline bioinformatique STARK

Le développement de STARK a suivi les recommandations émises par le groupe INCa pour la « conception de logiciels pour le diagnostic clinique par séquençage haut-débit » (http://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Conception-de-logiciels-pour-le-diagnostic-clinique-par-sequencage-haut-debit).

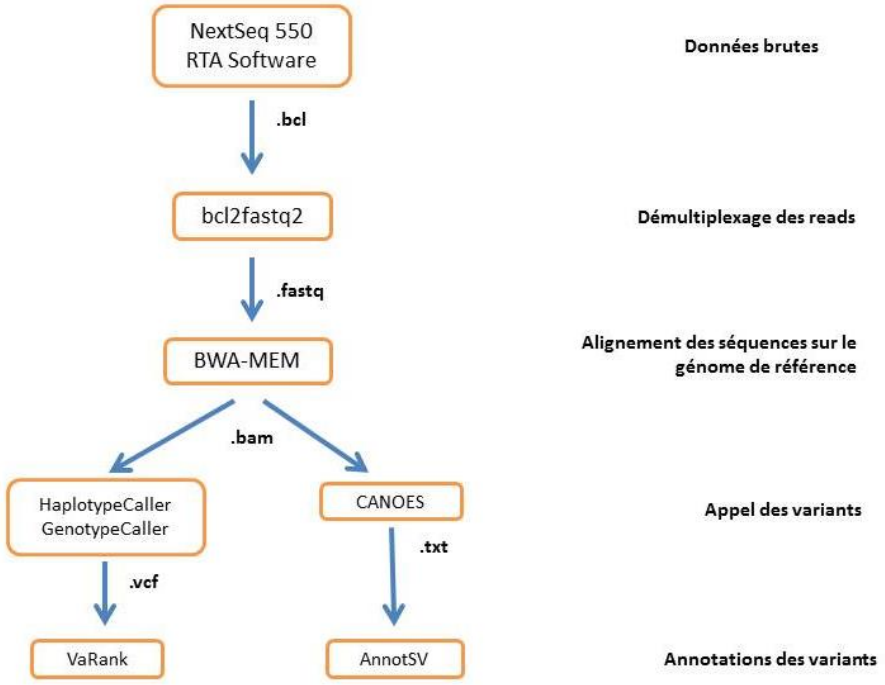
En outre, chaque version du code est tracée et maitrisée suivant un cycle de développement du logiciel et un environnement de travail distinct de la production (serveur HUX144DEV) est utilisé pour ce développement ainsi que pour les tests de validation.

3. Description du pipeline bioinformatique STARK

Le pipeline STARK comprend différents modules permettant de réaliser une chaine d'analyse (GENMOL_S5_MOPE_009_Bioinfo_Documentation_pipeline_STARK). Cette chaine d'analyse combine les outils les plus classiquement utilisés dans le contexte de la génétique constitutionnelle :

- Démultiplexage des reads (bcl2fastq2, Illumina)
- Analyse de la qualité des séquences et de l'alignement (FastQC)
- Alignement des reads sur le génome complet (BWA-MEM)
- Marquage des duplicats sans élimination des duplicats (MarkDuplicates)
- Ré-alignement des indels (GATK)
- Appel des variants sur les régions cibles (GATK HaplotypeCaller UnifiedGenotyper, CANOES)
- Recalibration des variants (GATK)
- Combinaison des VCFs de l'ensemble des pipelines en un VCF unique
- Annotation des variants (VaRank, AnnotSV)

L'ensemble des références des outils est décrit dans le fichier « env_tools.sh »



1. Programme d'interprétation des variants

Les annotations et les priorisations des variants identifiées par séquençage à haut débit sont réalisées au moyen du logiciel VaRank (Geoffroy V et al, 2015) intégré dans STARK. Les fichiers de variants sont par la suite analysés dans un tableur.
Ce logiciel permet de collecter les annotations des variations sur différentes bases de données biologiques utiles à l'interprétation telles que les (1000Genomes, ExAC, Clinvar …), sur des logiciels de prédiction avec un impact sur la protéine (SIFT, PolyPhen-2, …) et/ou sur des logiciels de prédiction d'effet sur l'épissage (NNSplice, MaxEntScan, …).

2. Banque interne de variants

Les variants identifiés pour chacun des patients séquencés ainsi que leur fréquence au sein de la cohorte pour chacun des secteurs sont stockés sous format *VCF* (fichiers de détection des variants du pipeline bioinformatique). Ces informations sont mises à disposition du biologiste au moyen des fichiers de résultats du logiciel VaRank.

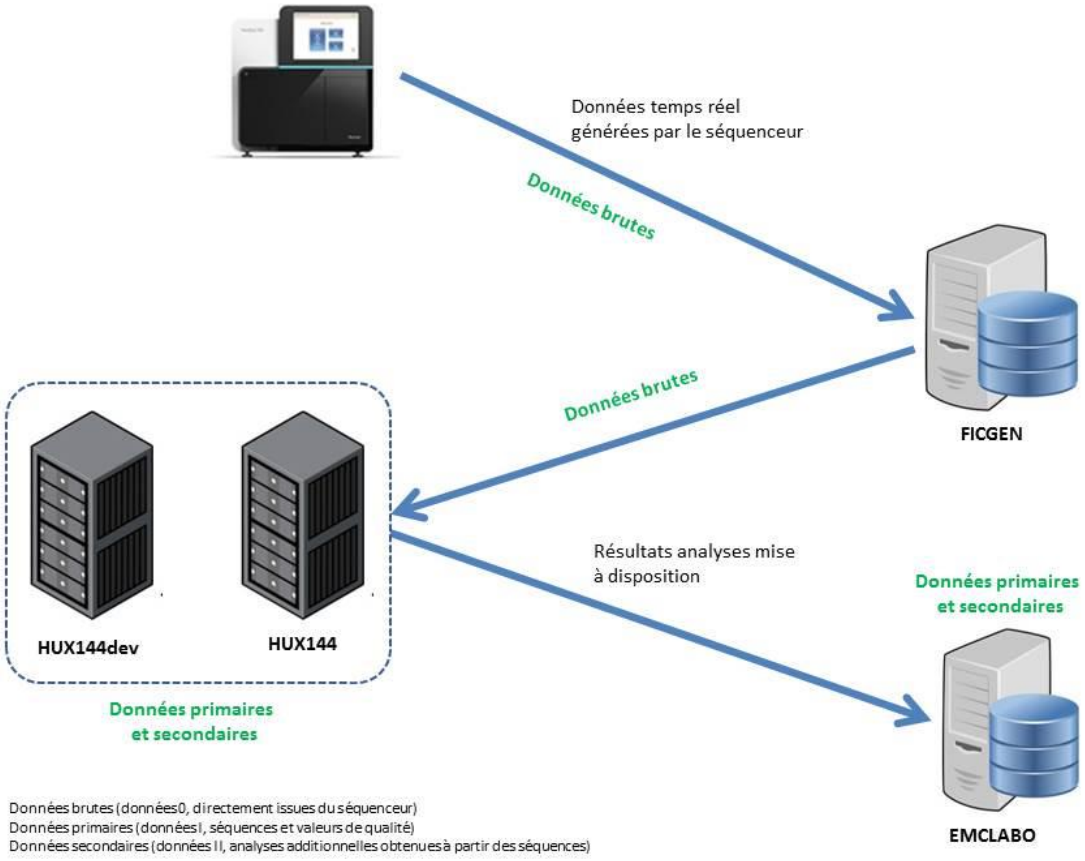
3. Programme de gestion des automatisations

STARK intègre une surveillance du répertoire (sur le serveur FICGEN) de dépôt des données brutes du séquenceur à haut débit. Ceci permet, dès le dépôt et la complétude des critères requis (ex : « RTAComplete.txt » et « SampleSheet.csv »), le lancement des analyses bioinformatiques.
(cf GENMOL_S5_MOPE_001_Bioinfo_Creation_de_la_SampleSheet)

4. Choix des ressources (équipements) et de l'environnement informatique en condition de production

Deux serveurs de calcul sont mis à disposition par le CRIH pour un usage exclusif aux analyses liées au séquençage à haut débit. Un serveur (HUX144) est utilisé en production et un serveur (HUX144dev) est utilisé pour les développements et les tests des nouvelles versions de STARK.
Deux serveurs de stockage sont également mis à disposition, FICGEN permet de stocker les données brutes issues du séquenceur et EMCLABO permet de stocker et mettre à disposition des utilisateurs les données primaires et secondaires.
(Cf. GENMOL_S5_MOPE_005_Bioinfo_Architecture_et_connexions_aux_serveurs_NGS)

5. Etablissement du schéma de circulation des données et des droits d'accès



Toutes les versions du code source du pipeline STARK sont conservées sur le serveur HUX144 et une copie est également réalisée sur EMCLABO.

Les données brutes du séquenceur sont déposées directement sur le serveur FICGEN accessible en lecture seule par le serveur HUX144. Ceci garantit la non-modification des données brutes. Ces données brutes sont ensuite traitées par STARK en données primaires et secondaires.
Les données primaires (ex : BAM) et secondaires (ex : VCF) sont transférées de manière sécurisée sur le serveur EMCLABO afin que les biologistes puissent y accéder et y réaliser leurs analyses. Une copie de ces données est conservée sur le serveur HUX144.

La gestion des utilisateurs et des droits d'accès aux serveurs est réalisée par le CRIH.
Sur FICGEN, seul le Nextseq550 a les droits en écriture.
Sur EMCLABO, les utilisateurs standards ont accès en lecture aux répertoires d'analyses (EMC_LABO\Archives\DIAG) et les bioinformaticiens ont les droits de lecture et d'écriture.

Un suivi des droits et utilisateurs sur les serveurs est réalisé HUX144, HUX144dev, FICGEN et EMCLABO\DIAG est effectué tous les 6 mois. (cf.GENMOL_S5_ENRG_002_Suivi_droits_et _utilisateurs_serveurs)

6. Définition des fichiers à stocker, à archiver et des durées de stockage/archivage

Les données brutes du séquenceur Illumina (ex : images) sont stockées sur FICGEN.
Les données primaires et secondaires sont stockées sur HUX144 de manière temporaire (~10 runs) et sur EMCLABO de manière pérenne.
Le CRIH procède à des sauvegardes en parallèle des données sur EMCLABO.

7. Analyse de risque

Données d'entrée	Points critiques à maîtriser (facteurs susceptibles d'influencer le résultat)	Criticité (1 non critique – 3 très critique)	Modalités de maîtrise	Documents de référence associés (à renseigner obligatoirement)
	Main d'œuvre			
Bioinformaticiens	Compétence du personnel	2	Diplôme, formation, expérience	Dossier RH
Bioinformaticiens	Compétence du personnel	1	Formation interne	GENMOL_S5_MOPE_009_Bioinfo_Documentation_pipeline_STARK DGEN-S5-GENMOL_S5_MOPE_001_Bioinfo_Creation_de_la_SampleSheet GENMOL_S5_MOPE_005_Bioinfo_Architecture_et_connexions_aux_serveurs_NGS GENMOL-S5-MOPE-003_Gestion_IGV_DGEN GENMOL-S5-MOPE-002_Gestion_Alamut Visual
Bioinformaticiens	Continuité de service	1	Planning	Planning Diagnostic génétique
	Milieu			
/	/	/	/	/
	Matériel			
Serveurs	Intégrité transmission des données entre HUX144 et EMCLABO	1	La commande « rsync » utilisée inclue un contrôle d'intégrité de type checksum md5. Boucle continue tant que toutes les données ne sont pas transférées.	Guide d'utilisation
	Méthode			

Données d'entrée	**Points critiques à maîtriser (facteurs susceptibles d'influencer le résultat)**	**Criticité (1 non critique – 3 très critique)**	**Modalités de maîtrise**	**Documents de référence associés (à renseigner obligatoirement)**
Pipeline STARK	Sauvegarde du pipeline	1	Code source de STARK archivé sur EMCLABO	Sur le serveur \\emclabo\archives_laboratoires\Archives\NGS\TOOLS
Pipeline STARK	Base de données interne (VaRank)	1	Reconstruction automatique de la base de données à partir des données stockées	Cf Document de VaRank (/home1/TOOLS/tools/varank/Guide_utilisateur_VaRank.pdf)
Pipeline STARK	Sauvegarde des données secondaires	1	Données disponibles sur deux serveurs indépendants (HUX144, EMCLABO) : - Sauvegarde des données sur le serveur de stockage (EMCLABO) - Conservation des données sur le serveur de calcul (HUX144)	GENMOL_S5_MOPE_005_Bioinfo_Architecture_et_connexions_aux_serveurs_NGS
Pipeline STARK	Gestion des dysfonctionnements du serveur de calcul	1	Procédure dégradée serveur : Utilisation du serveur de développement	GENMOL_S5_MOPE_009_Bioinfo_Documentation_pipeline_STARK
Pipeline STARK	Gestion des dysfonctionnements de l'application STARK	1	Procédure dégradée application : Utilisation de la version n-1 de STARK	GENMOL_S5_MOPE_009_Bioinfo_Documentation_pipeline_STARK
Pipeline STARK	Suivi des versions	1	Suivi des versions dans HUX144 et reportés dans le rapport et les fichiers de configuration Analyse d'un jeu de données : serveur de développement, serveur de production sans pointage, serveur de production avec pointage Modification et validation des versions n+1 sur un serveur d'intégration Test de non régression : set SNV/indel, set CNV et sets run réels	Cf. Guide d'utilisation de STARK Elément de preuve : rapport avec la bonne version de STARK GENMOL_S5_MOPE_004_Procédure_de_Validation_et_mise_en_production_de_STARK Communication par mail de la nouvelle version et information en réunion de service
Séquence de référence	Suivi des versions	1	Hg19 : fichier en lecture seule, non modifiable	Dossier /home1/TOOLS/genomes/hg19 sur HUX144
Bases de données et logiciels tiers de STARK	Suivi des versions	1	Liste des bases de données utilisées Traçabilité des versions utilisées dans le pipeline	Fichier env_tools.sh dans l'application STARK
	Matières premières			
/	/	/	/	/

8. Qualification du pipeline STARK pour la détection de SNV et d'insertion/délétion de petite taille à l'état homozygote ou hétérozygote

Le pipeline bioinformatique STARK doit être capable de détecter 100 % des variations SNV et indel de moins de 20 bases présentes dans un jeu de donnée validé et doit donner un génotype (type et position de variations, statut homozygote ou hétérozygote) correct.
La performance du pipeline STARK en **version 0.9.15b** a été évaluée sur un jeu de données de validation (données issues de matériel biologique et les mutations ont été confirmées par méthode Sanger).

a. Description de la méthode

Analyse de trois jeux de données externes dont les mutations présentes dans le vcf fourni ont été validées par Sanger (vcf_confirmé). Seules certaines variations retenues par le laboratoire ayant généré le jeu de données ont été mise à disposition. De même la zone validée en séquençage Sanger n'a pas été indiquée ce qui nous a empêché de déterminer la spécificité.
STARK a été appliqué aux 3 jeux de données (à partir des données primaires) et le résultat est présenté ci-dessous. Ces données ont été mises à disposition de la communauté par l'INCa (cf documents de Qualification_Matériau_Référence_Bioinfo pour chacun des jeux de données).

- BORDEAUX
  - 21 échantillons
  - Gènes étudiés *BRCA1, BRCA2, MLH1, MSH2, PTCH1* et *PTEN*
  - +/-50pb des régions ciblées
  - Un total de 17 SNVs et 11 Indels sont attendus

- ROUEN
  - 15 échantillons
  - Gènes étudiés *MSH6, MUTYH, STK11, MLH1, TP53, PTEN* et *APC*
  - +/-50pb des régions ciblées
  - Un total de 10 SNVs et 11 Indels sont attendus

- CPSGEN
  - 28 échantillons
  - Gènes étudiés *BRCA1* et *BRCA2*
  - +/-50pb des régions ciblées
  - Un total de 384 SNVs et 53 Indels sont attendus

b. Critères d'acceptabilité retenus

Le critère d'acceptabilité est la détection de 100% des variations présentes dans les jeux de données.

c. Résultats des expériences

L'analyse STARK a été réalisée à partir des fichiers *fastq* et en utilisant l'environnement « env.DIAG.sh ».
Pour le calcul de la spécificité de détection, un filtre sur la profondeur de lecture a été appliqué afin de ne retenir que les variant couvert à plus de 30X. Ce calcul a été effectué à partir des fichiers « vcf_confimé » fournis avec les jeux de données. (cf. GENMOL_S5_MOPE_004_Procédure_de_Validation_et _mise_en_production_de_STARK).

- BORDEAUX
  - 27 variants attendus ont été retrouvées par le pipeline STARK
  - 1 variant de type insertion, sur le gène *PTEN* (chr10 89720646 T TTT), n'a pas été détecté pour l'échantillon « GC-BOR-MIS-CAP-PTEN-4H42». C'est une région polyT en fin de l'exon 8 qui est une région difficile pour l'alignement et la détections des variants mais STARK a tout de même détecté à une position proche un variant de type « chr10 89720633 C CT » mais non concluant.

- ROUEN : Les 21 variants attendus ont été retrouvés par le pipeline STARK.

- CPSGEN : Les 437 variants attendus ont été retrouvées par le pipeline STARK.

d. Conclusion

En conclusion, STARK a une sensibilité :

- >99,25% pour la détection des SNV avec une confiance de 95% : critère atteint
- égale à 98,6% pour la détection des indels : critère atteint (cf argumentaire)

9. Qualification du pipeline STARK pour la détection des CNV

Le pipeline bioinformatique STARK est capable de détecter des CNV à l'aide de l'outil « CANOES » (Backenroth D et al, 2014).

a. Description de la méthode

Analyse de jeux de données interne et externe dont les CNV ont été validées par qPCR. Ces données ont été fournis par les biologistes du laboratoire afin d'implémenter notre set de validation des CNV et ainsi tester la capacité de notre pipeline à détecter les CNV connus via l'outil « CANOES ».

Un set de 10 jeux de données est actuellement utilisé pour la validation (cf documents de Qualification_Matériau_Référence_Bioinfo_CNV).

- BBS_1
  - 2 CNV sont attendus sur les gènes *ARL6* et *BBS1*

- BBS_2
  - 4 CNV sont attendus sur les gènes *TTC8, BBS9* et *BBS1*

- BBS_3
  - 1 CNV est attendu sur le gène *BBS4*

- BBS_AHC5LTAFXX
  - 1 CNV est attendu sur le gène *BBS4*

- CPS_AF206
  - 4 CNV sont attendus sur le gène *BRCA1* et *BRCA2*

- CPS_AF2LB
  - 1 CNV est attendu sur le gène *PALB2*

- DI_IGBMC_1
  - 1 CNV est attendu sur le gène *PTCHD1*

- MYOPATHIE_160928_NB551027_0056_AHGNJ5AFXX
  - 2 CNV sont attendus sur le gène *RYR1*

- MYOPATHIE_IGBMC
  - 1 CNV est attendus sur le gène *ISP*

b. Critères d'acceptabilité retenus

Le critère d'acceptabilité est la détection de 100% des CNV présentes dans le jeu de données.

c. Résultats des expériences

STARK a été appliqué sur l'ensemble du jeu de données. (cf. GENMOL_S5_MOPE_004_Procédure_de_Validation_et _mise_en_production_de_STARK)

- BBS_1
  - Les 2 CNV attendus ont été retrouvés par « CANOES »
- BBS_2
  - Les 4 CNV attendus ont été retrouvés par « CANOES »
- BBS_3
  - Le CNV attendu a été retrouvé par « CANOES »
- BBS_AHC5LTAFXX
  - Le CNV attendu a été retrouvé par « CANOES »
- CPS_AF206
  - 3 CNV attendus ont été retrouvés par « CANOES » et le quatrième est une délétion de 126pb qui a été retrouvé par STARK dans le final.vcf
- CPS_AF2LB
  - Le CNV attendu a été retrouvé par « CANOES »
- DI_IGBMC_1
  - Le CNV attendu a été retrouvé par « CANOES »
- MYOPATHIE_160928_NB551027_0056_AHGNJ5AFXX
  - 1 CNV attendus a été retrouvés par « CANOES » et le deuxième est une délétion de 91pb qui a été retrouvé par STARK dans le final.vcf
- MYOPATHIE_IGBMC
  - Le CNV attendu a été retrouvé par « CANOES »

d. Conclusion

En conclusion, CANOES a une sensibilité >82,5% avec une confiance de 95%. Le critère d'acceptabilité est atteint.

10. Analyse de la reproductibilité du pipeline STARK

La reproductibilité du pipeline a été testée sur un échantillon analysé à 5 reprises.
Une concordance de 100% a été obtenue sur les variations détectées.

11. Limite de détection du pipeline STARK

Le pipeline STARK et son environnement env.DIAG.sh ont été mis au point afin de détecter des variations :

- avec une profondeur de lecture minimale à une position génomique donnée de 30X.
- constitutionnelles

12. Incertitude de mesure

NA

13. Intervalle de référence/valeurs seuils

NA

14. Conclusion et suivi des performances du pipeline STARK

Conclusion

Le pipeline est qualifié pour la détection des SNV et petites insertions délétions pour des régions présentant une couverture suffisante (> 30X ou lectures) et pour la détection des CNV.

Suivi des performances

Un test de non régression est réalisé à chaque changement de version, ou au minimum, tous les 6 mois. (Cf. GENMOL_S5_MOPE_004_Procédure_de_Validation_et_mise_en_production_de_STARK)

15. Bibliographie

Backenroth D, Homsy J, Murillo LR, Glessner J, Lin E, Brueckner M, Lifton R, Goldmuntz E, Chung WK, Shen Y. CANOES: Detecting rare copy number variants from whole exome sequencing data. Nucleic Acids Research 2014)

Geoffroy V.*, Pizot C.*, Redin C., Piton A., Vasli N., Stoetzel C., Blavier A., Laporte J. and Muller J. VaRank: a simple and powerful tool for ranking genetic variants. PeerJ. 2015

DECLARATION d'APTITUDE
Conclusion : pipeline conforme Autorisé par : Jean MULLER, bioinformaticien, le 22/06/2018 Signature