Guide pour la maitrise d'un pipeline bioinformatique pour le séquençage haut débit

Date de rédaction : 24/10/2016	Date de révision :	Version : 1.0

Ce document a été préparé par les membres du GT4 « qualité et séquençage haut débit »

Groupe de travail : Cécile Aquaviva (Lyon), Christine Bellane-Chantelot (Paris), Stéphane Beziau (Nantes), Eric Bieth (Toulouse), Nelly Burnichon (Paris), Laurent Castera (Caen), Bénédicte Gérard (Strasbourg), Claude Houdayer (Paris), Eulalie Lasseaux (Bordeaux), Antony Le Bechec (Strasbourg), Adrien Pagin (Lille), Jean Muller (Strasbourg), Nicolas Sevenet (Bordeaux), Cécile Saint-Martin (Paris), Dominique Vaur (Caen)

Table des matières

Préambule

Un pipeline bioinformatique correspond à une chaine de traitement informatisée de données biologiques (séquences, variants). Dans cette application, le pipeline bioinformatique doit permettre l'identification de variation de séquence d'un échantillon par rapport à une séquence de référence.
Le champ d'application du pipeline bioinformatique doit être précisé (détection des SNV, taille des insertions ou délétions, détection des CNV, réarrangements type inversion, translocation, détection de séquence Alu …), ainsi que les limites de détection (par exemple, détection des mutations constitutionnelles, détection des mutations en mosaïque > 5 % …).

*Les **faux négatifs** bioinformatiques (absence de report d'une mutation pourtant présente dans les échantillons, test de sensibilité) peuvent avoir différentes causes :*

- *Elimination de reads par les filtres automatiques (trop de mésappariements dans le fragment et élimination des bam par trimming, couverture insuffisante et élimination dans le vcf, biais de brin, % de mutation trop faible et élimination dans le vcf …), clipping des mutations en fin d'amplicons*

Non élimination de fragments présentant une hybridation sur différents endroits du génome

*Les **faux positifs** (présence dans le fichier de sortie du pipeline d'une mutation qui n'existe pas dans les échantillons, test de spécificité) peuvent avoir différentes causes. Par exemple :*

- *Défaut de trimming (mauvaise élimination des amorces ou des données du séquenceur de mauvaise qualité…)*
- *Non élimination de fragments présentant un multimap (lecture alignée simultanément à plusieurs positions du génome)*
- *Défaut de réalignement local des indel*
- *Défaut d'annotation altérant significativement la signification du variant (par exemple allèles complexes type AA>TG rapportés séparément A>T et A>G et ont des prédictions totalement différentes sur la protéine)*

Les pipelines utilisés doivent pouvoir s'adapter au mieux à la balance souhaitée entre la sensibilité (faux négatif) et la spécificité (faux positif) nécessaire à l'interprétation des résultats.

1 Présentation des logiciels et du workflow bioinformatique

Le séquenceur génère des données brutes (images de fluorescence, variation de pH) traduites en base par le logiciel associé au séquenceur (bases et valeur de qualité associée) et constituent les données primaires (données I, séquences et valeurs de qualité). Les logiciels fournis par le fournisseur peuvent permettre également de faire l'étape de démultiplexage (capacité de séparation des données de séquence des différents échantillons analysés en parallèle).

Le pipeline bioinformatique doit permettre, à partir des données primaires (avant ou après démultiplexing) la détection de la présence d'une variation (ou « variant ») qualitative de séquence nucléotidique par rapport à une séquence de référence (e.g. génome de référence, identifié par une version « hg » ou « GRCh ») puis l'identification (position par rapport à la séquence de référence) et la qualification (e.g. qualité d'identification, profondeur, génotype homozygote/hémizygote ou hétérozygote, fréquence allélique) de cette variation. Ces données sont considérées comme les données secondaires (données II, analyses additionnelles obtenues à partir des séquences).

D'autres analyses bioinformatiques complémentaires permettent ensuite d'interpréter la variation en termes de pathogénicité.

<u>Donnée d'entrée :</u> données brutes de séquence de qualité suffisante

- Filtrage des données brutes
  - Marquage/suppression des adaptateurs et primers de séquence
  - Marquage/suppression des bases de mauvaise qualité (trimming)
- Alignement sur la séquence de référence
- Post-alignement sur la séquence de référence
  - Marquage/suppression des read dupliqués (si capture)
  - Réalignement local des indels (optionnel)
  - Recalibration de la qualité des bases
- Identification des variants (variant calling)
- Post-identification des variants
  - Recalibration des variants

- Annotation des variants
    - Filtrage des données qui seront interprétées (e.g. profondeur >30X, Valeur de qualité >50)
    - Priorisation/Ranking des variants (optionnel)
- Métriques
    - Evaluation du taux de couverture des séquences ciblées
    - Evaluation de la répartition des variants (e.g. nombre de polymorphismes, variants exoniques/introniques)

Donnée de sortie : fichiers de variants filtrés (.vcf, et éventuellement .csv, tsv, txt ou xls…) et parfois rankés.

Autant que possible, les étapes incluant un filtre doivent être non destructives, c'est-à-dire qu'il est privilégié de marquer les données (bases, reads, variants) plutôt que de les supprimer.

Maitrise des logiciels et du pipeline d'analyse

1.1 Logiciels et bases de données

Il est recommandé de vérifier périodiquement les nouvelles versions des logiciels disponibles (e.g. GATK, BWA, SAMtools, cutAdapt, picard) ou de bases de données (e.g. génome de référence, OMIM, ClinVar, ExAC, NNS, séquences de référence (« NM_XXXXXX »)).

Les numéros de version des logiciels et des bases de données doivent être tracés. *(Par exemple, dans les rapports des patients, le numéro de version du pipeline ou un enregistrement traçant les versions des logiciels utilisés)*

1.2 Description du workflow d'analyse bioinformatique

L'enchainement des analyses et les paramètres, notamment utilisés pour le filtrage de données (e.g. QV insuffisantes, données redondantes, trimming) doivent être décrits.

Recommandations générales selon Gargis *et al.*, Nature Rev Biotech 2015, 33 (7): 689-693 Good laboratory practice for clinical next-generation sequencing informatics pipelines

- *Elimination de l'analyse des reads dont les index ont des mésappariements*
- *Utilisation d'index avec au moins deux bases d'écart pour un démultiplexage non ambiguë*
- *L'alignement doit se faire de façon préférentielle sur le génome entier pour éviter les alignements forcés et évaluer le taux de « off target » (séquence hors de la cible de séquençage initiale)*

En cas d'utilisation d'un logiciel commercial, Il est conseillé d'utiliser les paramètres recommandés du logiciel et d'effectuer des tests de validation appropriés à chaque changement de paramètre.

2 Eléments de maitrise du pipeline bioinformatique

La maitrise du pipeline bioinformatique se base sur une analyse détaillée du 5M du pipeline et sur les éléments de maitrise

Données d'entrée	Points critiques à maîtriser	*Modalités de maîtrise exemples*
Main d'œuvre (habilitation du personnel) : **Préciser les références des procédures et enregistrements.**	Habilitation du personnel technique	*Fiches habilitation deux niveaux : utilisateur, développeur* *MOPE* *documentation des commandes propres à chaque logiciel* *lancement du pipeline* *création d'une samplesheet* *procédure de débuggage* *procédure dégradée…*
Milieu :	NA	
Matériaux de références (témoins) :	Echantillons de référence	*Dossier de qualification initiale du pipeline (tests)* *Par exemple,* *Analyse d'un set de données primaires (fastq ou .bam pour life) issu d'un ADN de référence ou d'un autre laboratoire accrédité ; comparaison des variants identifiés (sur les données ayant une couverture par exemple >30X)* *Analyse de sets de données issus d'échantillons ayant été préalablement analysés par la méthode de référence (séquençage Sanger).*
	Contrôles de qualité (CQI)	*Mise en place d'un contrôle interne de référence* *Analyse périodique d'un jeu de données restreint pour vérifier la performance du pipeline informatique.* *Suivi continu de la performance du pipeline bioinformatique (fichier de log, absence de fuite de données, reprise sur erreur…)*
Matériel :	Serveur	*Enregistrement de la version de l'OS*
	Dimensionnement du serveur de calcul et du réseau par rapport aux flux de données	*Tests de performance du pipeline*
	Sauvegarde	*Modalité de sauvegarde des logiciels informatiques utilisés et du pipeline bioinformatique proprement dit* *Tests périodiques de vérification de l'intégrité du stockage informatique*
	Sécurité	*Hébergeur agréé de données de santé (externalisation)*
	Logiciels d'analyse et dépendances	*Tests de vérification en cas de modification de N° de version de logiciels ou bases de données du pipeline bioinformatique* *Enregistrement des numéros de version des logiciels et bases de données utilisés, date d'application* *Enregistrement du pipeline bioinformatique (N° version, date d'application)* *Cahier de vie du pipeline bioinformatique*

2.1 Qualification initiale du pipeline bioinformatique

Un pipeline doit être validé avant toute utilisation. Les preuves de maitrise sont fournis soit par le prestataire interne, soit par le prestataire externe lorsque le traitement bioinformatique est réalisé par un fournisseur extérieur.

La qualification initiale du pipeline bioinformatique peut se faire grâce à un set de données artificielles (fastQ et vcf) ou à un set de données générées par une plateforme accréditée, utilisant si possible la même technologie, ou un set de données « gold standard ». Elle peut se faire également par l'analyse des données de séquence issus d'échantillons ayant été préalablement génotypés par une méthode de référence (séquençage Sanger).

Si un test de validation global du logiciel doit être réalisé, chaque étape peut néanmoins faire l'objet de tests spécifiques

Etapes du pipeline	***Exemples de tests***
Etape de démultiplexage	*Test de fidélité de démultiplexage* *Test de compatibilité des index (> ou = 2 bases de différences)* *Test de non contamination (intra- et inter-run)*
Etape d'élimination des adaptateurs et primers de séquence	
Etape d'élimination des bases de mauvaise qualité (trimming)	
Etape d'alignement sur le génome	*Vérification du numéro de version du génome téléchargé ??*
Etape d'élimination ou marquage des read dupliqués (si capture)	
Etape de variant calling	*Test utilisant différentes logiciels de variant caller en fonction des mutations retrouvées dans l'application clinique souhaité* *Test d'optimisation du variant caller à réaliser sur set de données artificielles ou/et réelle* *Test utilisant différents logiciels d'alignements pour voir les différences d'identification des variants*
Etape de réalignement local des indel (optionnel)	
Etape d'annotation des variants	
Etape de ranking des variants (optionnel)	*Test de vérification de la priorisation des variants sur des sets de données de patients.*
Etape d'évaluation du taux de couverture des séquences ciblées	*Vérification sur qq gènes cibles de la correspondance entre le chiffre de couverture calculée en sortie de pipeline et la couverture observée sur les .bam, sur IGV par exemple.*
Performance du pipeline bioinformatique	*Test de performance du pipeline (temps machine nécessaire et identification de variants ; à réaliser sur des sets de données réelles)*

2.2 Suivi du pipeline bioinformatique

Toute modification du pipeline doit être tracée. Une validation par des tests appropriés est nécessaire pour toute modification impactant le fonctionnement global du pipeline. Les outils utilisés pour cette validation doivent être le plus précisément identifiés (e.g. version, paramètres).

Par exemple, lorsqu'une nouvelle version de logiciel ou de base de données est implémentée dans le pipeline bioinformatique, des tests appropriés doivent être réalisés avant mise en production de la nouvelle version du pipeline bioinformatique pour vérifier que la performance du nouveau pipeline (e.g. rapidité, taux d'identification correct des variants, sensibilité/spécificité, non régression) n'est pas altérée.

La validation des mises à jour peut se faire, par exemple, par la réanalyse d'un même set de données primaires (fastq ou bam) et la comparaison des variants identifiés et/ou des données de couverture en fonction des modifications apportées.

2.3 Indicateur de suivi de la performance du pipeline bioinformatique

Au moins un indicateur doit être suivi pour suivre la performance opérationnelle du pipeline après *identification des paramètres critiques du paramétrage :*

Par exemple,

- *nombre total de variant*
- *% de variant connus (>90 %) en précisant la base de données de référence*
- *% d'indel (typiquement entre 10-15 %)*
- *% de variants avec statut homozygote*
- *% de variant non-sens (entre 0 à 5 si WES)*
- *% de transition/transversion (2,8 si WES, 2 si WGS)*
- *Analyse d'un fichier FastQ témoin à chaque run et vérification des couvertures et de la présence du ou des variants d'intérêt*

Une anomalie sur l'indicateur peut résulter d'un dysfonctionnement du pipeline ou des procédures analytiques en amont.

Définir la conduite en cas d'anomalie observée sur l'indicateur.