Qualification des solutions bioinformatiques-note technique		
Date de création : 25/02/2019	Date de révision :	Version : 1
Date de 1ère Application : 01/8/2019	N° document : NGSDIAG_005	
Approbation par le board du réseau le 01/08/2019		

Qualification des solutions bioinformatiques: note technique

Composition du groupe de travail :
Sylvie Bannwarth (Biologiste, CHU Nice), Pierre Blanc (Biologiste, AURAGEN ; Expert Technique Cofrac), Agnès Bourillon (Ingénieure technique/qualité, APHP), Nadège Calmels (Biologiste, CHU Strasbourg), Laurent Castera (Biologiste, Centre François Baclesse), Florence Coulet (Biologiste, APHP), Marie De Tayrac (Biologiste, CHU Rennes), Claude Houdayer (Biologiste, CHU Rouen), Marie Karam (Ingénieur qualité, CHU Nantes), Eulalie Lasseaux (Biologiste, CHU Bordeaux), Philippe Lochu (Biologiste, Genbio ; Evaluateur Technique Cofrac), Delphine Mallet-Motak (Ingénieur technique, CHU Lyon), Jean Muller (Bioinformaticien, CHU Strasbourg), Jean-François Taly (Bioinformaticien, Eurofins), Nancy Uhrhammer (Biologiste, Centre Jean Perrin),

Correspondance : pierre.blanc2@chu-lyon.fr

Abréviations utilisées dans ce document :
CIL : Comparaison inter-laboratoire
CIQ : Control interne de la qualité
EEQ : Évaluation Externe de la Qualité
LBM : Laboratoire de Biologie Moléculaire

Sommaire :

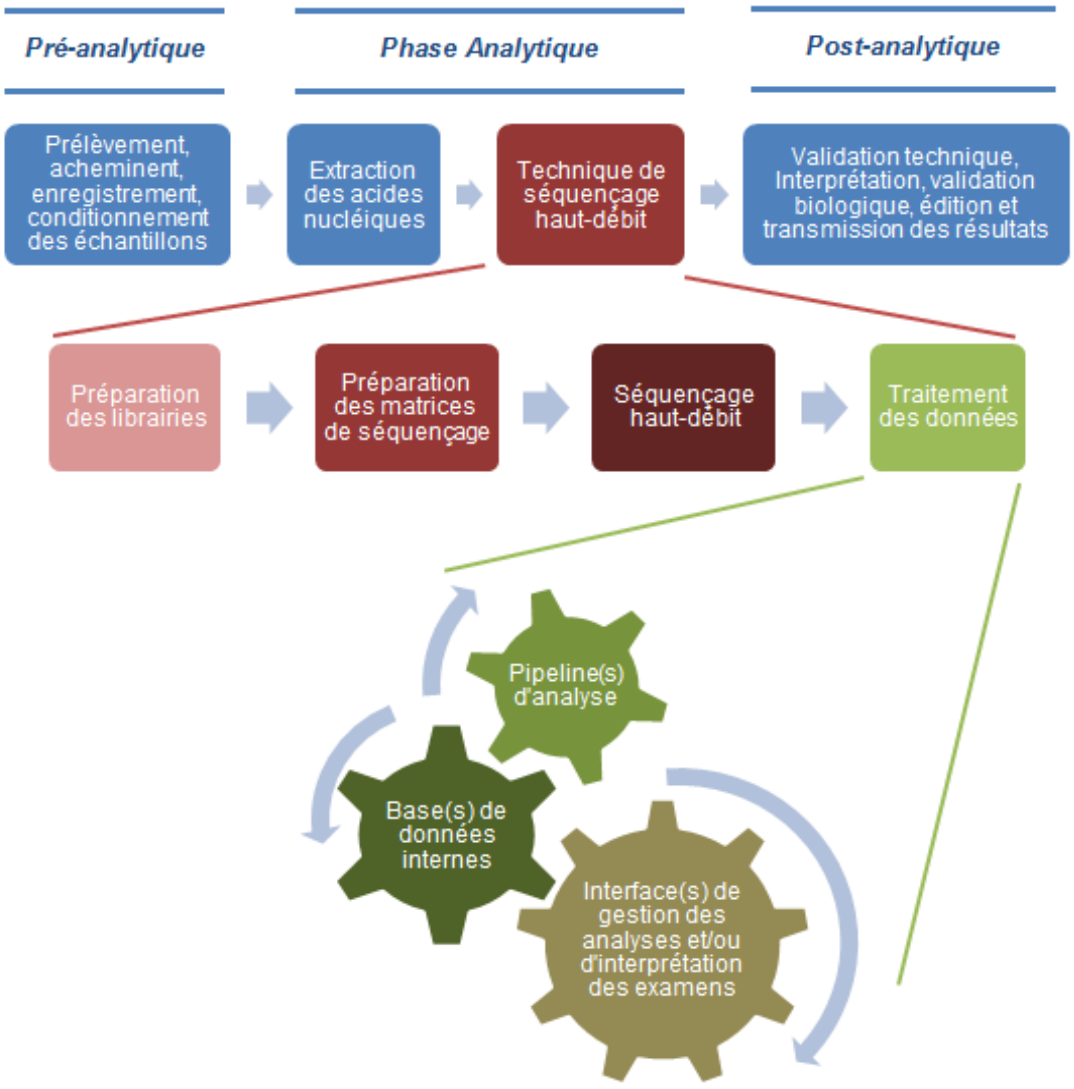
Avertissements

La présente **note technique**, ainsi que **le dossier de qualification** et l'**analyse des risques** fournis en exemples, ne constituent pas des recommandations. Les laboratoires de biologie moléculaire (LBM) sont libres de s'en inspirer ou de procéder différemment.

Comment appréhender les outils bioinformatiques dans la démarche d'accréditation ?

Le cœur de métier d'un LBM n'est pas de développer des logiciels selon des standards industriels. Le séquençage haut-débit nécessite néanmoins la mise en œuvre de **chaines de traitement des données ou pipelines** adaptés aux plateformes analytiques, aux besoins des utilisateurs et à la finalité des examens. *De facto*, un certain nombre de LBM ont été amenés à développer un ou plusieurs pipelines en interne.

A ces pipelines, peuvent s'ajouter d'autres outils de traitement : **bases internes de métriques qualité et de variants** (voir recommandations 15 et 21 EuroGentest-ESGG, PMID : 26508566), **interface graphique d'interprétation des variants**, voire surcouche informatique de **gestion et d'automatisation du flux des données** entre ces différentes instances.



Le développement initial puis continu de ces outils logiciels ainsi que leur maintenance requiert des ressources humaines et matérielles qu'il peut être difficile d'obtenir puis de pérenniser ou d'encadrer au sein d'un LBM. Cette contrainte a conduit un certain nombre de laboratoires à externaliser, de manière partielle ou totale, ces outils de traitements des données.

A l'intermédiaire entre ces deux cas de figures, les solutions de traitement de données peuvent être embarquées sur les séquenceurs ou installées sur un serveur associé.

Le traitement bioinformatique des données issues du séquençage haut-débit fait partie intégrante de la **phase analytique**.

Que la bioinformatique soit développée en interne, « embarquée » ou externalisée, le biologiste responsable du LBM demeure responsable de cette étape analytique, au même titre que de l'ensemble des phases de l'examen (voir recommandation 2 de AMP-CAP-AMIA, PMID : 29154853).

En pratique, il est souhaitable qu'un biologiste -qualifié au séquençage haut-débit et à l'interprétation des données correspondantes- soit impliqué dans :

- l'expression du cahier des charges et le développement des outils (le cas échéant),
- la qualification des outils,
- la validation de méthode globale.

Les outils logiciels bioinformatiques peuvent être considérés à la fois comme des équipements du laboratoire (matériels d'analyse) et des systèmes de traitement de données informatiques (respectivement au 5.3.1 et au 5.10.3 de la Norme NF EN ISO 15189). A ce titre, ils doivent faire l'objet d'une **qualification** ou **essai d'acceptation**.

L'**essai d'acceptation** consiste à vérifier que les outils logiciels atteignent les **spécifications et performances attendues** dans les **conditions de production du laboratoire**.

Quand réaliser la qualification d'une solution bioinformatique ?

Préalablement à leur qualification dans l'environnement de production du LBM, les solutions bioinformatiques auront fait l'objet d'une **phase de développement**, soit en dehors du laboratoire (fournisseur de séquenceurs, prestataire spécialisé…), soit au sein du laboratoire (service de génétique moléculaire, service support…).

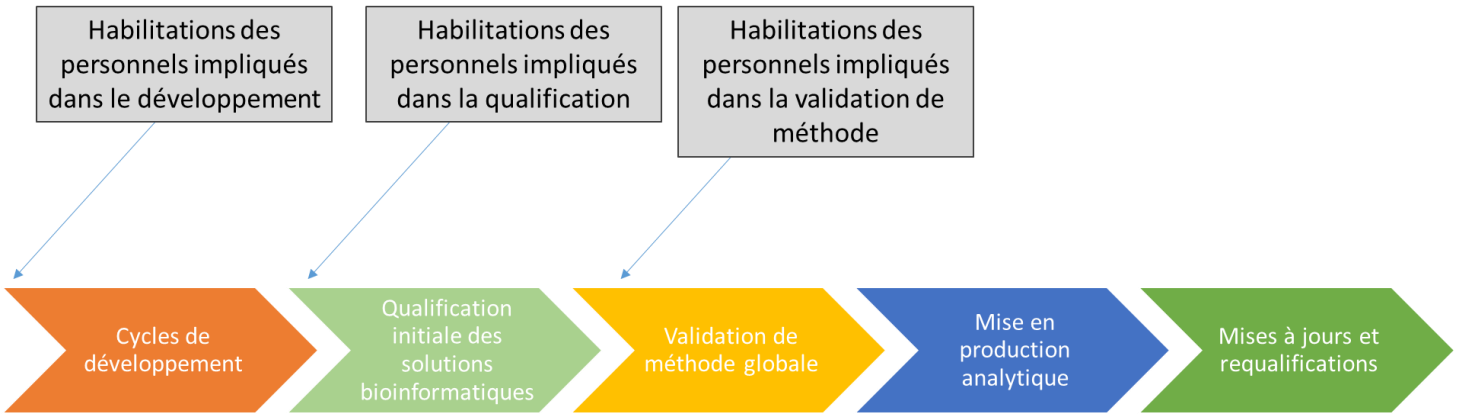
Dans ce dernier cas, il est conseillé de se reporter aux recommandations suivantes :

- Conception de logiciels pour le diagnostic clinique par séquençage haut-débit, INCa, février 2018. ISBN 978-2-37219-362-7ISBN, Net 978-2-37219-363-4 http://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Conception-de-logiciels-pour-le-diagnostic-clinique-par-sequencage-haut-debit

- Recommandations conjointes de l'*Association of Molecular Pathology*, du *College of American Pathologists* et de l'*American Medical Informatics Association* (PMID : 29154853).

Ces recommandations peuvent également aider les laboratoires à établir leurs **critères de sélection** et d'**évaluation** des fournisseurs bioinformatiques.
Comme pour tout équipement intervenant dans la phase analytique, la qualification des outils bioinformatiques est réalisée **préalablement à la validation de méthode globale**. L'essai d'acceptation est alors référencé dans le dossier de validation de méthode correspondant.

Enfin, le personnel du laboratoire impliqué dans la qualification doit être **habilité** à cette tache (voir page 8) **au préalable de la qualification**.



Comment qualifier un outil bioinformatique développé en interne ?

Pipelines, bases de données, interfaces de visualisation et d'interprétation… les systèmes bioinformatiques participent au moins à l'une des fonctions énumérées au **5.10.3 de la Norme NF ISO 15189** : collecte, traitement, enregistrement, stockage, récupération des données analytiques. Comme précisé dans la même section, les **fonctions informatiques** de ces systèmes doivent être **vérifiées avant application**. Le laboratoire définit les **fonctions informatiques** qu'il souhaite vérifier, ainsi que ses **critères d'acceptation**.

L'essai d'acceptation peut porter sur l'**ensemble de l'environnement bioinformatique** et/ou sur ses **composants** (pipelines, bases, etc.).

Les tests sont, en principe, dirigés par l'**analyse des risques** réalisée pendant la phase de développement.

Plusieurs initiatives existent ou se développent pour accompagner les laboratoires dans l'**analyse des risques (initiale et dynamique)**, ainsi que dans la mise en œuvre **d'actions préventives et correctives.** A titre d'exemples :

- Répertoire d'articles sur les problèmes rencontrés en séquençage haut-débit : https://sequencing.qcfail.com/

- Projet d'inventaire des erreurs de séquençage (EMQN), suite à la session 'erreurs bioinformatiques' de l'ESHG Meeting 2018 à Milan : https://emqn.us12.list-manage.com/track/click?u=b5507ff445e1fedcc9ca57599&id=f4e112741c&e=aedeabdf42

Les tests réalisés lors de la qualification peuvent être de plusieurs types :

- **Tests unitaires :** vérification de la fonctionnalité de chacun des modules d'un pipeline (i.e. contrôle des fichiers intermédiaires d'un pipeline)
- **Tests fonctionnels :** vérification de chaque exigence du cahier des charges, par exemples :
  - temps d'exécution d'un pipeline,
  - vérifications des sorties attendues sur un fichier CIQ (fichier de patient connu, données *in silico*…)
  - gestions des variants multi-alléliques par une interface de visualisation,
  - etc.
- **Tests de robustesse :** vérification de la stabilité et de la fiabilité dans le temps, par exemples :
  - comportement d'un pipeline sur fichier corrompu (ex : erreur lors d'un transfert ou d'une compression) ou bien sur fichier erroné (ex : erreur typographique dans une '*sample sheet*'),
  - reprise sur erreur (procédure dégradée et retour sur point) ou après défaillance système (exemple : coupure de réseau)
  - modulation de la volumétrie des données et résistance au pic de charge,
  - traçabilité des erreurs logicielles ou matérielles (logs)
  - stabilité d'une base de variants ou de métriques après sauvegarde,
  - stabilité de l'annotation des variants dans une interface de visualisation des résultats,
  - etc.
- **Test de sécurité :** vérification de l'intégrité (absence de fuite de données) et de la confidentialité des données

Tout comme les matrices utilisées pour l'essai d'acceptation d'un séquenceur (Phi X, librairies issus d'échantillons biologiques), ces tests peuvent être conduits avec des données de complexité variable. Il existe plusieurs types de **fichiers 'étalon'** (FASTQ, BAM, VCF…):

- *In silico*
- internes au laboratoire, dont le contenu aura été préalablement caractérisé (fichier CIQ)
- issus de CIL
- issus d'EEQ
- de référence, reposant sur des intervalles génomiques et collections de variants à haute confiance (voir notamment PMID 30936564)

Enfin, il peut être utile pour les outils de type **pipelines**, d'adosser des **tests de performance** aux tests de qualification proprement dits. Notamment, lorsqu'un pipeline est utilisé pour l'analyse de plusieurs panels de gènes, les **performances génériques** du pipeline peuvent être évaluées à l'occasion de la qualification (comme c'est le cas dans le dossier joint en exemple).

Les tests de performances sont basés sur des **comparaisons de VCF**. En la matière, il est conseillé de suivre les recommandations suivantes:

- *Global Alliance for Genomic and Health* (https://www.ga4gh.org);
- *PrecisionFDA* (https://precision.fda.gov/docs/comparisons);
- PMID 30858580, voir notamment *supplementary information* : https://static-content.springer.com/esm/art%3A10.1038%2Fs41587-019-0054-x/MediaObjects/41587_2019_54_MOESM1_ESM.pdf

Les jeux de données contenant des **intervalles et variants à 'haute confiance'** permettent d'extraire de la comparaison les positions : vraies positives (**VP**), vraies négatives (**VN**), fausses positives (**FP**) et fausses négatives (**FN**).

Sont alors principalement calculés :

- **La sensibilité analytique bioinformatique: VP / (VP + FN)**
  Il s'agit d'une mesure de l'**exhaustivité** du pipeline.
  Nb : Sensibilité = '*recall*' or '*sensitivity*'

- **La valeur prédictive positive (VPP) analytique bioinformatique : VP / (VP + FP)**
  Il s'agit d'une mesure de l'**exactitude** du pipeline.
  Nb : Valeur prédictive positive = '*positive predictive value*' or '*precision*'

- **La F-mesure ('*F-score*') ou mesure F1: 2 * (sensibilité * VPP) / (sensibilité + VPP)**
  C'est la **moyenne harmonique** de la sensibilité et de la VPP, intégrée aux métriques de l'EMQN à compter de la campagne 2018-2019.

Comment qualifier un outil bioinformatique « embarqué » ou fourni par un prestataire externe ?

Dans un cas comme dans l'autre, il est conseillé de disposer de **compétences bioinformatiques minimales** pour **maitriser les outils logiciels** et être capable d'**identifier les principales causes d'altération des résultats** (exemple : paramétrage de la détection des variants ou « *variant calling* »).

Solution(s) embarquée(s) sur un séquenceur ou serveur associé :

- Idéalement, l'essai d'acceptation du séquenceur est couplé à celui des solutions bioinformatiques associées. Il est détaillé par le fournisseur, approuvé par le laboratoire et fait l'objet d'une **attestation de qualification** de la part du fournisseur.
- Alternativement, lorsque la qualification n'a pas été réalisée par le fournisseur à l'initial ou bien n'est plus envisageable *a posteriori*, le laboratoire peut être amené à qualifier lui-même les solutions bioinformatiques.
- Pour ce faire, la méthode à fois la plus simple et la plus exhaustive peut consister à **qualifier l'ensemble de l'infrastructure bioinformatique (liaisons, serveurs et logiciels bioinformatiques) en point final**. Cela revient dans ce cas à utiliser un ou plusieurs fichiers étalon, à réaliser un test de performance et à s'assurer que la sensibilité et la VPP (*a minima*) atteignent les valeurs attendues par le laboratoire (voir supra).

Solution(s) fournie(s) par un prestataire externe :

- Les solutions proposées peuvent revêtir de nombreuses configurations :
  - Analyse bioinformatique totalement externalisée, accès aux résultats via une interface de visualisation et d'interprétation (exemple : *Sophia Genetics**)
  - Gamme de prestations flexibles (exemple : *SeqOne**)
- Il est conseillé aux laboratoires de distinguer la **qualification** des solutions proprement dite de l'**accompagnement éventuel à la validation de méthode globale**.
- La qualification d'une solution externalisée doit reposer sur un **niveau d'exigence comparable** à celui mise en œuvre dans le cadre d'une solution interne (voir paragraphe correspondant, page 4).
- Les compétences internes peuvent naturellement être un facteur limitant, mais il est néanmoins conseillé au laboratoire de s'assurer de la **transparence de l'essai d'acceptation** et de sa compréhension *a minima*.
- A cet effet, les **types de test** mis en œuvre lors de la qualification sont explicités par le fournisseur, revus et approuvés par le laboratoire.
- Les **résultats des tests** sont présentés par le fournisseur, revus et approuvés par le laboratoire.
- Le fournisseur délivre une **attestation de qualification**, dissociée de l'éventuel **rapport de validation de méthode**.
- Il est conseillé aux laboratoires de **contractualiser l'ensemble de leurs besoins avec le prestataire**, y compris les **pièces documentaires** requises pour constituer leur dossier de qualification (voir ci-dessous).

Comment construire le dossier de qualification initiale?

Il est conseillé de préciser dans le dossier de qualification (ou d'attacher en annexe) :

- **Cahier des charges du laboratoire** (résumé des principales exigences informatiques et biologiques)
- **Description synthétique de la solution bioinformatique** (exemple : schéma modulaire d'un pipeline)
- **Description de la méthode mise en œuvre lors de son développement**
- **Description de l'architecture bioinformatique et de la circulation des données** (exemple : un schéma peut résumer le flux des données au travers des réseaux et serveurs de production hébergeant les solutions bioinformatiques)
- **Modalités et droits d'accès aux données**
- **Modalités de versionnage et de mise à jour des solutions bioinformatiques**
- **Analyse des risques**
- **Certifications ISO (9001, 13485, 27001…), certification HDS, déclaration de conformité RGPD**
- **Méthode (dont description des échantillons 'étalon') et résultats des tests de qualification.** Les principaux résultats peuvent être rapportés dans le dossier. Par commodité, les résultats plus détaillés peuvent être gérés via un outil informatique (Gitlab par exemple) ou déposés dans un répertoire dédié, qui seront alors référencés dans le dossier.
- **Manuel utilisateur**
- **Personnels habilités ayant participé à la mise en œuvre et à la validation des tests.**

Comment aborder le changement et la requalification ?

Les solutions bioinformatiques et leurs composants évoluent sans doute plus rapidement encore que les protocoles techniques de production de séquence.

La gestion des mises à jour ainsi que des requalifications est donc un enjeu important.

Il appartient au laboratoire de définir les risques associés au changement.

Il est donc suggéré de faire appel aux principes d'**intégration continue** utilisés en génie logiciel. L'intégration continue est un ensemble de pratiques consistant à **vérifier à chaque modification de code source que le résultat des modifications ne produit pas de régression dans l'application développée**.

Selon l'importance ou la criticité des modifications, la requalification peut être abordée au moyen de:

- **Tests de non régression ciblés**
- **Tests de non régression globaux :** au maximum il peut s'agir d'un nouveau test de performance portant sur l'ensemble de l'infrastructure bioinformatique
- **Tests automatiques :** en cas de mises à jour programmées, sur CIQ

Exemples :

- Modification du module de traitement des données FFPE : requalification d'un 'pipeline somatique'
- Remplacement sur serveur de calcul, mise à jour annuel du système d'exploitation : requalification de l'ensemble des pipelines
- Mise à jour trimestrielle d'une base de données externe : requalification programmée

Les valeurs clefs retournées par les tests, les chemins vers les fichiers de référence et la périodicité des requalifications peuvent être codés dans les fichiers de configuration de chacun des outils bioinformatiques.

Ce faisant, il est possible de distinguer l'opérateur chargé du développement du pipeline et celui chargé de la qualification périodique. Le but est ici d'adapter le personnel aux compétences requises pour chacune de ces tâches.

Que les solutions bioinformatiques soient internes, « embarquées » ou bien externalisées, il appartient au laboratoire de s'assurer qu'il maitrise l'évolution de ses solutions bioinformatiques et qu'il peut en documenter le contenu.

Comment établir l'habilitation initiale et le maintien des compétences des bioinformaticiens du laboratoire ?

Le 5.1 de la Norme, relatif au personnel du laboratoire, constitue l'une des principales sources d'écarts. Un **bioinformaticien** appartenant au périmètre du laboratoire doit, comme tout autre personnel de ce laboratoire :

- Avoir des fonctions définies (§5.1.3) : fiche de fonction
- être accueilli dans son environnement de travail (§5.1.4)
- être formé à l'initial (§5.1.5) et en continu (§5.1.8)
- être habilité à l'initial (= qualifié) aux fonctions qu'il occupe (§5.1.2)
- être maintenu régulièrement dans ses compétences (§5.1.6 & 5.1.7)

Dans la pratique, la prise de fonction des bioinformaticiens peut avoir précédé de plusieurs mois, voire de plusieurs années, la qualification des outils logiciels et la demande d'extension d'un examen de séquençage haut-débit.
D'autre part les compétences requises pour prononcer l'habilitation à ce type de fonction très spécifique peuvent apparaitre comme un facteur limitant.

Les **modalités** d'habilitation initiale et de maintien des compétences d'un bioinformaticien posent donc fréquemment question.

Le recrutement d'un bioinformaticien s'appuie toutefois nécessairement sur la **sélection d'un profil de compétence** et s'accompagne de l'**attribution de missions et objectifs**. Ces éléments peuvent fournir le support des critères d'habilitation initiale.

Quelques exemples de critères d'**habilitation initiale** et de preuves :

- **Formations professionnelles:** diplômes ; CV signé par le bioinformaticien et approuvé-signé par le biologiste responsable du laboratoire ; avis du personnel décisionnaire lors du tour de recrutement, validé-signé par le biologiste responsable…
- **Formation initiale (attention aux exigences du 5.1.5):** attestation de participation aux formations, résultats des tests post-formation (exemple : formation à la Norme NF EN ISO 15189, au logiciel de gestion documentaire…)
- **Formation continue:** attestation de présence aux congrès, journées réseau/société savantes et aux formations métier de l'année antérieure…
- **Implication effective dans le développement des outils:** impressions d'écran (dans Gitlab ou autre), certificat de mission du biologiste responsable ou du chef de service…
- **Participation à l'amélioration continue des solutions logicielles pendant la phase de développement, d'optimisation et de familiarisation :** impressions d'échanges mails, tickets, requêtes Redmine…

Le **maintien des compétences** peut être dérivé des trois derniers points (formation continue, développement & amélioration continue des solutions), auxquels peut être ajoutée la participation au traitement des CIQ, CIL et EEQ.

Références

Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, Race V, Sistermans E, Sturm M, Weiss M, Yntema H, Bakker E, Scheffer H, Bauer P; EuroGentest; European Society of Human Genetics (2016) **Guidelines for diagnostic next-generation sequencing.** *Eur J Hum Genet* 24(1):2-5. PMID: 26508566

Conception de logiciels pour le diagnostic clinique par séquençage haut-débit, collection Outils pour la pratique, INCa, février 2018. ISBN 978-2-37219-362-7ISBN, Net 978-2-37219-363-4

Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV, Wang C, Carter AB1 (2018) **Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists**. *J Mol Diagn*. 20(1):4-27. PMID: 29154853

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M (2014) **Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.** *Nat Biotechnol*. 32(3):246-51. PMID: 24531798

Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, De La Vega FM, Xiao C, Sherry S, Salit M (2019) **An open resource for accurately benchmarking small variant and reference calls**. *Nat Biotechnol*. [Epub ahead of print] PMID: 30936564

Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, Lababidi S, Truty R, Asimenos G, Funke B, Fleharty M, Chapman BA, Salit M, Zook JM; Global Alliance for Genomics and Health Benchmarking Team (2019) **Best practices for benchmarking germline small-variant calls in human genomes**. *Nat Biotechnol*. [Epub ahead of print] PMID : 30858580