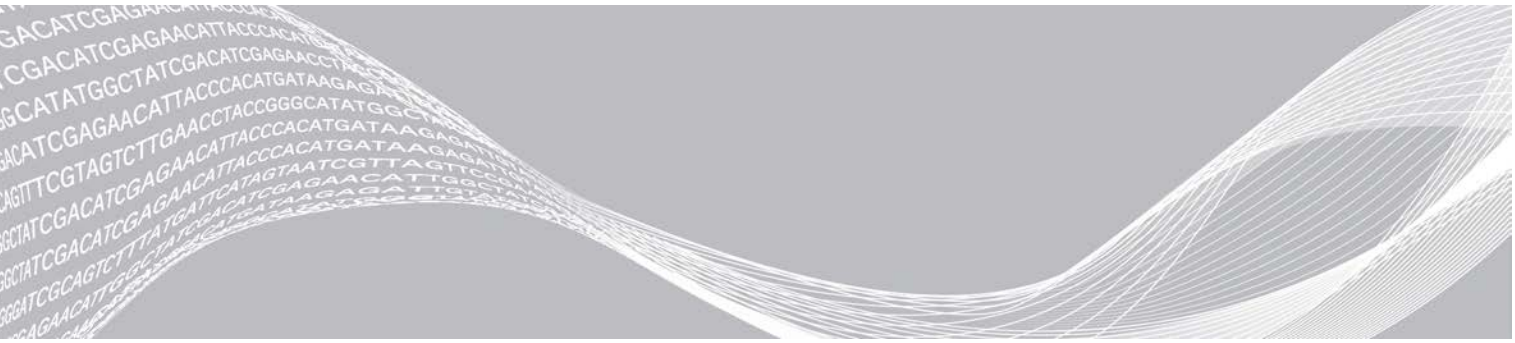
illumina®

Platforme DRAGEN Bio-IT v3.5 d'Illumina

Guide de l'utilisateur



Document n° 1000000111887 v00 FRA

EXCLUSIF À ILLUMINA

Février 2020

Destiné à la recherche uniquement.
Ne pas utiliser dans le cadre d'examens diagnostiques.

Historique des révisions

Document	Date	Description des modifications
Document n° 1000000111887 v00	Février 2020	Ajout des options suivantes : • --intermediate-results-dir • --vc-enable-liquid-tumor-mode • --vc-tin-contam-tolerance • --vc-enable-af-filter • --vc-enable-non-homref-normal-filter Ajout des valeurs par défaut pour les options de sous-échantillonnage de la définition de petits variants. Ajout des renseignements sur le filtrage de fichier BAM d'entrée. Modification de --vc-tlod-call-threshold à --vc-sq-call-threshold. Modification de --vc-tlod-filter-threshold à --vc-sq-filter-threshold. Mise à jour des filtres de définition postsomatique disponibles et des filtres postsomatiques de contrôle de référence correspondants. Mise à jour des options multiéchantillons de l'analyse combinée. Ajout des renseignements sur les régions d'élimination pour l'étape des dénombrements de cibles. Ajout du champ de génotype de l'entrée FORMAT du fichier VCF de variants du nombre de copies. Ajout des renseignements sur la fonction de définition de variants somatiques du nombre de copies de la plateforme DRAGEN. Replacement des renseignements sur la fonction de définition de variants structurels Manta par les renseignements sur la fonction de définition de variants structurels de la plateforme DRAGEN. Ajout des renseignements sur la fonction d'estimation de la ploïdie. Ajout de l'indicateur de définition de variants Percent QC Region Callability in Region (Pourcentage de définissabilité de région de QC dans la région). Suppression du rapport sur la prédiction de la ploïdie. Suppression des renseignements sur le module de réétalonnage des scores de qualité de variants (vQSR). Ajout des renseignements sur le rapport de biais GC. Ajout des renseignements sur les identifiants moléculaires uniques (IMU). Suppression des options suivantes : • --vc-clustered-events-threshold • --vc-clustered-events-filter • --sv-reference • --sv-quiet

Table des matières

Chapitre 1 Plateforme DRAGEN Bio-IT d'Illumina

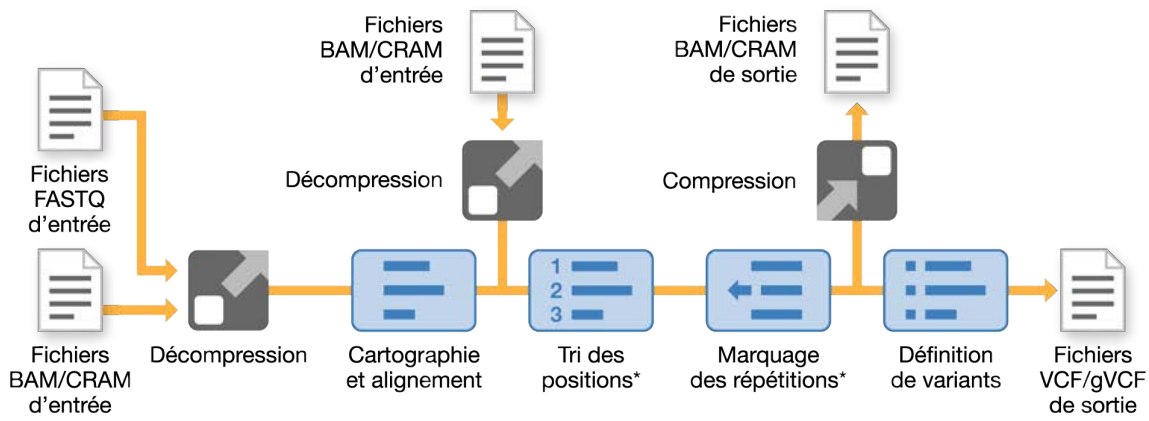
Le cœur de la plateforme DRAGENMC Bio-IT d'Illumina est le processeur hautement reconfigurable DRAGEN Bio-IT qui est intégré à une carte FPGA (Field Programmable Gate Array) et offert dans un serveur préconfiguré pouvant être parfaitement intégré dans des flux de travail bioinformatiques. Des algorithmes hautement optimisés peuvent être chargés dans la plateforme pour répondre à différents pipelines d'analyse secondaire de séquençage nouvelle génération (SNG). Ces pipelines comprennent :

- Génome entier
- Exome
- Séquençage de l'ARN
- Méthylome
- Cancer

Toutes les interactions des utilisateurs sont accomplies par l'intermédiaire du logiciel de la plateforme DRAGEN exécutant le serveur hôte et gérant toutes les communications avec la carte DRAGEN. Ce guide de l'utilisateur résume les aspects techniques du système et fournit des renseignements détaillés sur toutes les options de lignes de commande de la plateforme DRAGEN. Si vous travaillez avec le système DRAGEN pour la première fois, Illumina vous recommande de lire d'abord le *Guide d'introduction de la plateforme DRAGEN Bio-IT d'Illumina (1000000076675)* qui peut être téléchargé sur la page d'assistance du site Web d'Illumina (sous « Support »). Ce guide propose une introduction à la plateforme DRAGEN et comprend l'exécution d'un test du serveur, la génération d'un génome de référence et l'exécution d'exemples de commandes.

Pipeline d'ADN de la plateforme DRAGEN

Figure 1 Pipeline d'ADN de la plateforme DRAGEN



* Facultatif

Le pipeline d'ADN de la plateforme DRAGEN accélère grandement l'analyse secondaire des données de séquençage nouvelle génération (SNG). Par exemple, le temps nécessaire pour traiter un génome humain entier à une couverture de 30x est réduit d'environ 10 heures (en utilisant le logiciel standard actuel de l'industrie, BWA-MEM+GATK-HC) à environ 20 minutes. La durée varie en fonction de la profondeur de couverture.

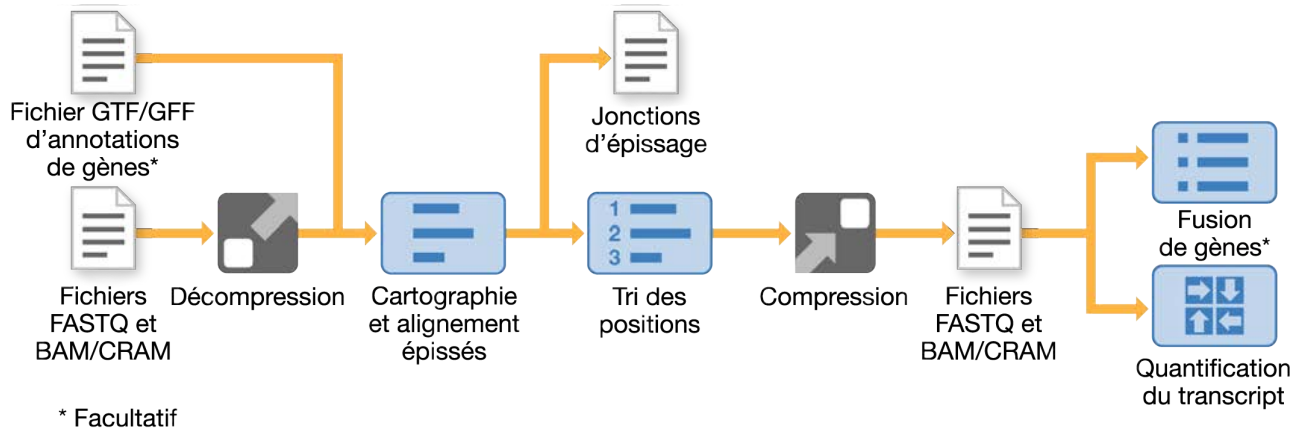
Ce pipeline exploite la puissance extraordinaire de la plateforme DRAGEN Bio-IT et comprend des algorithmes hautement optimisés pour la cartographie, l'alignement, le tri, le marquage des répétitions et la définition de variants haplotypes. Il utilise aussi des fonctionnalités de la plateforme comme la compression avec accélération matérielle et la conversion en format BCL optimisée, ainsi que l'ensemble complet des outils de la plateforme.

Contrairement à toutes les autres méthodes d'analyse secondaire, les applications du pipeline d'ADN de la plateforme DRAGEN ne sacrifient pas l'exactitude pour obtenir une exécution plus rapide. L'exactitude de la plateforme DRAGEN est meilleure que celle du logiciel BWA-MEM+GATK-HC, à la fois pour les polymorphismes mononucléotidiques et les indels (dans des comparaisons côte à côte).

En plus de la définition de variants haplotypes, le pipeline prend en charge la définition de variants du nombre de copies et de variants structurels, de même que la détection des expansions de répétition.

Pipeline d'ARN de la plateforme DRAGEN

La plateforme DRAGEN comprend une fonction d'alignement de séquençage de l'ARN (sensible à l'épissage), ainsi que des composants d'analyse propres à l'ARN pour la quantification d'expression génique et la détection de fusions de gènes.



Le pipeline d'ARN de la plateforme DRAGEN partage plusieurs de ses composants avec le pipeline d'ADN. La cartographie de séquences de point d'ancrage court à partir de lectures de séquençage de l'ARN est similaire à la cartographie de lectures d'ADN. Les jonctions d'épissage (la jonction des exons non contigus dans les transcrits d'ARN) à proximité des points d'ancrage cartographiés sont détectées et incorporées à l'ensemble des alignements de lectures.

La plateforme DRAGEN utilise des algorithmes d'accélération matérielle pour cartographier et aligner les lectures de séquençage d'ARN plus rapidement et plus exactement que les outils logiciels populaires. Par exemple, elle permet d'aligner 100 millions de lectures appariées de séquençage d'ARN en environ 3 minutes. Lorsque des ensembles de données de référence simulés pour le séquençage de l'ARN, sa sensibilité et sa spécificité de jonction d'épissage sont inégalées.

Pipeline de méthylation de la plateforme DRAGEN

Le pipeline de méthylation de la plateforme DRAGEN fournit un soutien à l'automatisation du traitement de données de séquençage au bisulfite en générant un fichier BAM contenant les étiquettes nécessaires à l'analyse de la méthylation.

Mises à jour systèmes

La plateforme DRAGEN est flexible, extensible et hautement reconfigurable. Votre abonnement DRAGEN vous permet de télécharger des mises à jour pour les processeurs et le logiciel de la plateforme DRAGEN. Ces mises à jour visent l'amélioration de la vitesse, des performances, du débit et de l'exactitude de la plateforme.

Ressources supplémentaires et assistance

Pour obtenir des renseignements et des ressources supplémentaires, des mises à jour système et de l'assistance, veuillez visiter la page d'assistance de la plateforme DRAGEN sur le site Web d'Illumina (sous « Support »).

Pour commencer

La plateforme DRAGEN vous fournit des tests à exécuter pour vous assurer que votre système est bien installé et bien configuré. Avant d'exécuter les tests, assurez-vous que le serveur DRAGEN est suffisamment alimenté et refroidi et qu'il est connecté à un réseau assez rapide pour transmettre vos données en provenance et à destination de la machine avec une performance adéquate.

Vérification du système

Après avoir mis le serveur en marche, vous pouvez vous assurer que le serveur DRAGEN fonctionne correctement en exécutant la commande /opt/edico/self_test/self_test.sh qui effectue les actions suivantes :

- Indexation automatique du chromosome M à partir du génome de référence hg19.
- Chargement du génome de référence et de l'index.
- Cartographie et alignement d'un ensemble de lectures.
- Enregistrement des lectures alignées dans un fichier BAM.
- Confirmation que les alignements correspondent exactement aux résultats attendus.

Chaque serveur envoie des données FASTQ de test pour ce script qui se trouvent sous /opt/edico/self_test. La vérification du système prend de 25 à 30 minutes.

L'exemple suivant montre comment exécuter le script ainsi que le résultat d'un test réussi :

```
[root@edico2 ~]# /opt/edico/self_test/self_test.sh
   --------------------------------
   test hash creating
   test hash created
   --------------------------------
   reference loading /opt/edico/self_test/ref_data/chrM/hg19_chrM
   reference loaded
   --------------------------------

   real0m0.640s
   user0m0.047s
   sys0m0.604s
   not properly paired and unmapped input records percentages: PASS
   --------------------------------
   md5sum check dbam sorted: PASS
```

```
--------------------------------
SELF TEST COMPLETED
SELF TEST RESULT : PASS
```

Si le fichier BAM de sortie ne correspond pas aux résultats attendus, alors la dernière ligne de texte contiendra plutôt :

```
SELF TEST RESULT : FAIL
```

Si vous recevez une mention « FAIL » (échec) après avoir exécuté le script de test immédiatement après la mise en marche de votre serveur DRAGEN, communiquez avec l'assistance technique d'Illumina.

Exécution de votre propre test

Une fois que les performances de votre système DRAGEN vous ont satisfait, vous êtes prêt à analyser vos propres données à l'aide de la machine, comme suit :

- Chargez la table de référence pour le génome de référence.
- Déterminez l'emplacement des fichiers d'entrée et de sortie.
- Traitez des données d'entrée.

Chargement du génome de référence

Avant qu'un génome de référence puisse être utilisé par la plateforme DRAGEN, il doit être converti en format binaire personnalisé à partir du format FASTA pour pouvoir être utilisé par les composants matériels du système DRAGEN. Pour en savoir plus, consultez la section *Préparation d'un génome de référence* à la page 141.

La table de hachage de référence spécifiée dans la ligne de commande est automatiquement chargée dans la carte la première fois que vous traitez des données à l'aide d'un pipeline. Vous pouvez charger manuellement la table de hachage de votre génome de référence à l'aide de la commande suivante :

```
dragen -r <reference_hash-table_directory>
```

Assurez-vous que le répertoire de la table de hachage de référence se trouve sur le lecteur du système de fichiers rapide.

L'emplacement par défaut de la table de hachage du génome hg19 est le suivant :

```
/staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149
```

La commande permettant de charger le génome de référence hg19 à partir de l'emplacement par défaut est la suivante :

```
dragen -r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149
```

Cette commande permet de charger le génome de référence binaire dans la mémoire de la carte DRAGEN, où il est utilisé pour traiter autant d'ensembles de données d'entrée que vous voulez. Vous n'avez pas besoin de charger le génome de référence de nouveau sauf si le système est redémarré ou si vous devez passer à un génome de référence différent. Le chargement d'un génome de référence peut prendre jusqu'à une minute.

La plateforme DRAGEN vérifie si le génome de référence spécifié se trouve déjà dans la carte. S'il y est, alors le chargement du génome de référence est automatiquement sauté. Vous pouvez forcer le rechargement du même génome de référence à l'aide de l'option de ligne de commande *force-load-reference (-l)*.

La commande permettant de charger le génome de référence inscrit les versions du logiciel et du matériel dans les fichiers de sortie standards. Par exemple :

```
DRAGEN Host Software Version 01.001.035.01.00.30.6682 and
Bio-IT Processor Version 0x1001036
```

Une fois le génome de référence chargé, le message suivant est inscrit dans les fichiers de sortie standards :

```
DRAGEN finished normally
```

Emplacements des fichiers d'entrée et de sortie

La plateforme DRAGEN Bio-IT est très rapide, ce qui nécessite une bonne planification des emplacements des fichiers d'entrée et de sortie. Si les fichiers d'entrée ou de sortie se trouvent dans un système de fichiers lent, alors la performance globale du système est limitée par le débit de ce système de fichiers. Il est recommandé que les fichiers d'entrée et de sortie soient transmis directement en provenance ou en direction d'un système de stockage externe.

Le système DRAGEN est préconfiguré avec au moins un système de fichiers rapide qui consiste en un ensemble de disques SSD rapides regroupés en RAID-0 aux fins de performance. Ce système de fichiers est monté sous /staging. Ce nom a été choisi pour souligner le fait que cet emplacement a été conçu pour être vaste et rapide, mais qu'il n'est pas redondant. La défaillance de n'importe lequel des disques du système de fichiers entraîne la perte de l'ensemble des données qui y sont stockées.

Durant le traitement, la plateforme DRAGEN génère et lit des fichiers temporaires. Lorsque vous utilisez la plateforme DRAGEN, il est fortement recommandé de toujours diriger les fichiers temporaires vers les disques SSD rapides (où à l'emplacement /staging) en utilisant l'option *--intermediate-results-dir*. Si l'option *--intermediate-results-dir* n'est pas spécifiée, les fichiers temporaires seront stockés dans le répertoire --output-directory. Nous vous recommandons de transmettre les fichiers d'entrée et de sortie de la plateforme DRAGEN à l'aide d'un système de fichiers externe.

Traitement de vos données d'entrée

Pour analyser des données de fichier FASTQ, utilisez la commande *dragen*. Par exemple, la commande suivante peut être utilisée pour analyser un fichier FASTQ à paire de bases unique :

```
dragen \
   -r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
   -1 /staging/test/data/SRA056922.fastq \
   --output-directory /staging/test/output \
   --output-file-prefix SRA056922_dragen \
   --RGID DRAGEN_RGID \
   --RGSM DRAGEM_RGSM
```

Pour en savoir plus sur les options de lignes de commande, consultez la section *Logiciel hôte de la plateforme DRAGEN* à la page 6.

Chapitre 2 Logiciel hôte de la plateforme DRAGEN

Le logiciel hôte *DRAGEN* de la plateforme DRAGEN est utilisé pour construire et charger des génomes de référence, puis pour analyser les données de séquençage en décompressant les données, effectuant la cartographie, l'alignement, le tri, le marquage des répétitions avec retrait facultatif et la définition des variants.

Vous pouvez utiliser le logiciel en utilisant la commande *dragen*. Les options de ligne de commande sont décrites dans les sections qui suivent.

Les options de ligne de commande peuvent également être inscrites dans un fichier de configuration. Pour plus de renseignements sur les fichiers de configuration, consultez la section *Fichiers de configuration* à la page 15. Si une option est inscrite dans le fichier de configuration et spécifiée dans la ligne de commande, l'option de ligne de commande est utilisée plutôt que celle du fichier de configuration.

Options de ligne de commande

Le résumé de l'utilisation de la commande *dragen* est comme suit :

- Construction de la référence et de la table de hachage

```
dragen --build-hash-table true --ht-reference <REF_FASTA> \
   --output-directory <REF_DIRECTORY> [options]
```

- Exécution de la cartographie et de l'alignement et de la fonction de définition de variants *(*.fastq à *.vcf)*

```
dragen -r <REF_DIRECTORY> --output-directory <OUT_DIRECTORY> \
   --output-file-prefix <FILE_PREFIX> [options] -1 <FASTQ1> \
   [-2 <FASTQ2>] --RGID <RG0> --RGSM <SM0> --enable-variant-caller true
```

- Exécution de la cartographie et l'alignement *(*.fastq à *.bam)*

```
dragen -r <REF_DIRECTORY> --output-directory <OUT_DIRECTORY> \
   --output-file-prefix <FILE_PREFIX> [options] \
   -1 <FASTQ1> [-2 <FASTQ2>] \
   --RGID <RG0> --RGSM
```

- Exécution de la fonction de définition de variants *(*.bam à *.vcf)*

```
dragen -r <REF_DIRECTORY> --output-directory <OUT_DIRECTORY> \
   --output-file-prefix <FILE_PREFIX> [options] -b <BAM> \
   --enable-variant-caller true
```

- Exécution de la fonction de conversion de données BCL *(BCL à *.fastq)*

```
dragen --bcl-conversion-only true --bcl-input-directory <BCL_DIRECTORY> \
   --output-directory <OUT_DIRECTORY>
```

- Exécution de la cartographie et l'alignement de l'ARN *(*.fastq à *.bam)*

```
dragen -r <REF_DIRECTORY> --output-directory <OUT_DIRECTORY> \
   --output-file-prefix <FILE_PREFIX> [options] -1 <FASTQ1> \
   [-2 <FASTQ2>] --enable-rna true
```

Pour consulter une liste complète des options de ligne de commande, consultez la section *Options de ligne de commande* à la page 162.

Options relatives au génome de référence

Avant de pouvoir vous servir du système DRAGEN pour aligner des lectures, vous devez charger sur la carte PCIe un génome de référence et les tables de hachages qui y sont associées. Pour en savoir plus sur le prétraitement des fichiers FASTA d'un génome de référence pour les mettre en formats de référence binaire et de table de hachage natifs à la plateforme DRAGEN, consultez la section *Préparation d'un génome de référence* à la page 141. Vous devez également spécifier le répertoire qui contient la référence en format binaire et les tables de hachage à l'aide de l'option *-r* [ou *--ref-dir*]. Cet argument est toujours requis.

Vous pouvez charger le génome de référence et les tables de hachage dans la mémoire de la carte DRAGEN séparément des lectures en traitement comme suit :

```
dragen -r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149
```

Vous pouvez utiliser l'option *-l (--force-load-reference)* pour forcer le chargement du génome de référence même s'il est déjà chargé comme suit :

```
dragen -l -r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149
```

Le temps nécessaire au chargement du génome de référence dépend de la taille de la référence. Avec les paramètres recommandés habituels, le chargement prend de 30 à 60 secondes.

Modes de fonctionnement

La plateforme DRAGEN a deux modes de fonctionnement principaux :

- Cartographie et alignement
- Définition de variants

Le système DRAGEN est capable d'exécuter chaque mode indépendamment ou dans le cadre d'une solution de bout en bout. Le système DRAGEN vous permet également d'activer et de désactiver la compression, le tri, le marquage de répétitions et la compression tout au long du pipeline DRAGEN.

- Mode pipeline de bout en bout

  Pour exécuter le mode pipeline de bout en bout, mettez à l'option *--enable-variant-caller* la valeur « *true* » (activée) et fournissez en entrée des lectures non cartographiées en format *.fastq, *.bam ou *.cram.

  La plateforme DRAGEN effectue la décompression, la cartographie, l'alignement, le tri, le marquage facultatif de répétitions et envoie directement ces données en entrée à la fonction de définition de variants pour produire un fichier VCF. Dans ce mode, la plateforme DRAGEN exécute des étapes en parallèle tout au long du pipeline pour réduire de beaucoup la durée d'exécution globale.

- Mode de cartographie et d'alignement

  Le mode de cartographie et d'alignement est activé par défaut. Les fichiers d'entrée contiennent des lectures non cartographiées et peuvent être en format *.fastq, *.bam, ou *.cram. La plateforme DRAGEN produit un fichier BAM ou CRAM présentant des lectures alignées et triées. Pour marquer des lectures comme répétitions en même temps, mettez à l'option *--enable-duplicate-marking* la valeur « *true* » (activée).

- Mode de définition de variants

  Pour exécuter le mode de définition de variants, mettez à l'option *--enable-variant-caller* la valeur « true » (activée).

  Le fichier d'entrée contient des lectures alignées et triées au format BAM. La plateforme DRAGEN produit un fichier VCF. Si les lectures du fichier BAM sont déjà triées, vous pouvez sauter le tri en mettant à l'option *--enable-sort* la valeur « *false* » (désactivée).

Le mode de définition de variants ne permet pas de marquer comme répétition les lectures des fichiers BAM dans le pipeline DRAGEN avant d'effectuer la définition de variants. Utilisez plutôt le mode de fonctionnement de bout en bout pour pouvoir utiliser la fonctionnalité de marquage de répétitions.

- Données de séquençage de l'ARN

  Pour activer le traitement des données provenant du séquençage de l'ARN, mettez à l'option *--enable-rna* la valeur « *true* » (activée).

  La plateforme se sert de la fonction d'alignement épissé de l'ARN pendant l'étape de cartographie et d'alignement. Le processeur de la plateforme DRAGEN Bio-IT bascule de manière dynamique entre les modes de fonctionnement nécessaires.

- Données de méthylation du séquençage au bisulfite

  Pour activer le traitement des données de méthylation du séquençage au bisulfite, mettez à l'option *--enable-methylation-calling* la valeur « true » (activée). La plateforme DRAGEN automatise le traitement des données pour les protocoles de Lister et de Cokus (également appelés protocole directionnel et protocole non directionnel) qui produit un fichier BAM doté d'étiquettes compatibles avec Bismark.

  Vous pouvez également exécuter la plateforme DRAGEN dans un mode qui produit un fichier BAM pour chaque combinaison de lectures de références converties de C à T et de G à A. Pour activer ce mode de traitement, vous devez construire un ensemble de tables de hachage de référence à l'aide de l'option *--ht-methylated*, puis exécuter la commande *dragen* en utilisant la valeur de l'option *--methylation-protocol* appropriée.

Le reste de cette section fournit davantage de renseignements sur les options qui permettent de régler le pipeline DRAGEN plus finement.

Options relatives à la sortie

Les options de sortie de ligne de commande suivantes sont obligatoires :

- L'option *--output-directory <out_dir>* spécifie le répertoire de sortie pour les fichiers produits.
- L'option *--output-file-prefix <out_prefix>* spécifie le préfixe pour les fichiers de sortie. La plateforme DRAGEN ajoute l'extension de fichier appropriée à ce préfixe pour chaque fichier produit.
- L'option *-r [--ref-dir]* spécifie la table de hachage de référence.

Par souci de concision, les exemples qui suivent n'incluent pas ces options obligatoires.

Pour ce qui est de la cartographie et de l'alignement, le fichier de sortie est trié et compressé au format BAM par défaut avant l'enregistrement sur le disque. L'utilisateur peut spécifier le format de sortie pour l'étape de cartographie et d'alignement à l'aide de l'option *--output-format <SAM|BAM|CRAM>*. Si le fichier de sortie existe déjà, le logiciel affiche un message d'avertissement et s'arrête. Pour forcer l'écrasement du fichier de sortie existant, utilisez l'option *-f [--force]*.

Par exemple, les commandes suivantes produisent un fichier BAM compressé et forcent l'écrasement.

```
dragen ... -f
dragen ... -f --output-format sam
```

Pour générer un fichier BAM d'index en format BAI (extension de fichier .bai), mettez à l'option *--enable-bam-indexing* la valeur « true » (activée).

L'exemple suivant envoie sa sortie dans un fichier SAM et force l'écrasement :

```
dragen ... -f --output-format sam
```

L'exemple suivant produit un fichier CRAM de sortie et force l'écrasement :

```
dragen ... -f --output-format cram
```

La plateforme DRAGEN génère des étiquettes de différence de mésappariements (MD, pour mismatch difference), comme décrites dans la spécification du format BAM. Toutefois, comme générer ces chaînes entraîne un petit coût en termes de performance, cette option est désactivée par défaut. Pour générer des étiquettes MD, mettez à l'option *--generate-md-tags* la valeur « true » (activée).

Pour générer des étiquettes d'état d'alignement ZS:Z, mettez à l'option *--generate-zs-tags* la valeur « true » (activée). Ces étiquettes ne sont générées que dans l'alignement primaire et seulement quand une lecture a des alignements sous-optimaux qui se qualifient comme fichier de sortie secondaire (même si aucun d'entre eux n'a été envoyé en sortie parce que l'option *--Aligner.sec-aligns* avait la valeur « 0 »). Les valeurs d'étiquette valides sont :

- ZS:Z:R – Plusieurs alignements avec des scores similaires ont été trouvés
- ZS:Z:NM – Aucun alignement n'a été trouvé
- ZS:Z:QL – Un alignement a été trouvé, mais il se trouve sous le seuil de qualité

Pour générer des étiquettes SA:Z, mettez à l'option *--generate-sa-tags* la valeur « true » (activée), la valeur par défaut. Ces étiquettes fournissent des renseignements sur l'alignement (position, chaîne CIGAR, orientation) des groupes d'alignements supplémentaires. Ceux-ci sont utiles à la définition de variants structurels.

Options relatives aux fichiers d'entrée

Le système DRAGEN est en mesure de traiter des lectures dans les formats FASTQ ou BAM/CRAM. Si vos fichiers FASTQ d'entrée se terminent par .gz, la plateforme DRAGEN décompresse automatique les fichiers à l'aide de la décompression avec accélération matérielle.

Fichiers FASTQ d'entrée

Les fichiers FASTQ d'entrée peuvent être des fichiers de lectures à paire de bases unique et de lectures appariées, comme le montrent les exemples suivants.

- Lectures à paire de bases unique dans un fichier FASTQ *(option « -1 »)*

```
dragen -r <REF_DIR> -1 <fastq> --output-directory <OUT_DIR> \
   --output-file-prefix <OUTPUT_PREFIX> --RGID <RGID> --RGSM <RGSM>
```

- Lectures appariées dans deux fichiers FASTQ correspondants *(options « -1 » et « -2 »)*

```
dragen -r <REF_DIR> -1 <fastq1> -2 <fastq2> \
   --output-directory <OUT_DIR> --output-file-prefix <OUT_PREFIX> \
   --RGID <RGID> --RGSM <RGSM>
```

- Lectures appariées dans un seul fichier FASTQ imbriqué *(option « --interleaved (-i) »)*

```
dragen -r <REF_DIR> -1 <INTERLEAVED_FASTQ> -i \
   --RGID <RGID> --RGSM <RGSM>
```

Les échantillons du fichier FASTQ peuvent être divisés dans plusieurs fichiers pour limiter la taille des fichiers ou réduire la durée de leur génération. Les commandes BCL de la plateforme DRAGEN et bcl2fastq utilisent une convention de nommage commune :

```
<SampleID>_S<#>_<Lane>_<Read>_<segment#>.fastq.gz
```

Par exemple :

```
RDRS182520_S1_L001_R1_001.fastq.gz
RDRS182520_S1_L001_R1_002.fastq.gz
...
```

```
RDRS182520_S1_L001_R1_008.fastq.gz
```

La convention de nommage de fichiers est différente pour les instruments HiSeqX et NextSeq.

Ces fichiers n'ont pas besoin d'être concaténés pour être traités ensemble par la plateforme DRAGEN. Pour cartographier et aligner un échantillon, fournissez-le avec le premier fichier de la série (*-1 <FileName>_001.fastq*). La plateforme DRAGEN lit consécutivement tous les fichiers de segments de l'échantillon pour les séquences des deux fichiers FASTQ spécifiés à l'aide des options « -1 » et « -2 » pour les lectures appariées d'entrée, ainsi que pour les fichiers fastq.gz. Ce comportement peut être désactivé en mettant à l'option *--enable-auto-multifile* la valeur « false » (désactivée) dans la ligne de commande.

La plateforme DRAGEN peut également lire plusieurs fichiers dont le nom contient le nom d'un échantillon donné, ce qui peut servir à combiner des échantillons qui se trouvent dans différentes lignes BCL ou Flow Cell. Pour activer cette fonctionnalité, mettez à l'option *--combine-samples-by-name* la valeur « true » (activée).

Si la convention de nommage de fichier Casava 1.8 illustrée ci-dessus est utilisée pour les fichiers FASTQ spécifiés dans la ligne de commande et que d'autres fichiers dans le même répertoire ont le même nom d'échantillon, ces fichiers et tous leurs segments seront automatiquement traités. Veuillez prendre note que le nom d'échantillon, le numéro de lecture et l'extension de fichier doivent être identiques. Le code à barres et le numéro de ligne peuvent être différents.

Les fichiers d'entrée doivent se trouver dans un système de fichier rapide pour éviter de ralentir la performance du système.

Fichier fastq-list d'entrée

Pour fournir plusieurs fichiers FASTQ d'entrée, il est recommandé d'utiliser l'option *--fastq-list <nom du fichier CSV>* pour spécifier le nom du fichier CSV qui contient la liste des fichiers FASTQ plutôt que l'option *--combine-samples-by-name*. Par exemple :

```
dragen -r <ref_dir> --fastq-list <CSV_FILE> \
   --fastq-list-sample-id <Sample_ID> \
   --output-directory <OUT_DIR> --output-file-prefix <OUT_PREFIX>
```

L'utilisation d'un fichier CSV permet de donner des noms arbitraires aux fichiers FASTQ d'entrée, d'utiliser des fichiers se trouvant dans différents sous-répertoires et de spécifier explicitement des étiquettes BAM pour chaque groupe de lectures. La plateforme DRAGEN génère automatiquement un fichier CSV au format approprié lors de la conversion du format BCL à FASTQ. Ce fichier CSV est nommé fastq_list.csv et contient une entrée pour chaque fichier FASTQ et chaque paire de fichiers de lectures appariées produits pendant cette analyse.

Format de fichier FASTQ à valeurs séparées par des virgules

La première ligne du fichier à valeurs séparées par des virgules contient le titre de chaque colonne et est suivie par au moins une ligne de données. Toutes les lignes du fichier CSV doivent contenir le même nombre de valeurs séparées par des virgules. Elles ne doivent pas contenir de blancs ou de caractères étrangers.

Les titres de colonnes sont sensibles à la casse. Les titres de colonnes suivants sont requis :

- RGID – Groupe de lectures
- RGSM – Identifiant de l'échantillon
- RGLB – Librairie
- Lane – Ligne Flow Cell
- Read1File – Chemin d'accès complet vers un fichier d'entrée FASTQ valide

- Read2File – Chemin d'accès complet vers un fichier d'entrée FASTQ valide (requis pour un fichier de lectures appariées en entrée, à laisser vide dans le cas contraire)

Chaque fichier FASTQ inscrit dans la liste de valeurs séparées par des virgules peut n'y être inscrit qu'une seule fois. Toutes les valeurs de la colonne Read2File doivent être remplies et faire référence à des fichiers valides, ou être toutes vides.

Lors de la génération d'un fichier BAM à partir d'un fichier d'entrée fastq-list, un groupe de lectures est généré pour chaque valeur RGID unique. L'en-tête du fichier BAM contient des étiquettes RG pour les champs suivants :

- ID (de la colonne RGID)
- SM (de la colonne RGSM)
- LB (de la colonne RGLB)

D'autres étiquettes peuvent être spécifiées pour chaque groupe de lectures en ajoutant un titre de colonne composé de quatre caractères majuscules commençant par « RG ». Par exemple, pour ajouter une étiquette « PU » (unité de plateforme), ajoutez une colonne intitulée « RGPU », puis spécifiez la valeur de chaque groupe de lectures dans cette colonne. Tous les titres de colonnes doivent être uniques.

Un fichier fastq-list peut comprendre des fichiers à utiliser pour plus d'un échantillon. Si un fichier fastq-list contient une seule entrée RGSM unique, alors aucune option supplémentaire n'a besoin d'être spécifiée. La plateforme DRAGEN traitera tous les fichiers inscrits dans le fichier fastq-list. S'il y a plus d'une entrée RGSM unique dans un fichier fastq-list, l'une des deux options suivantes doit être spécifiée en plus de l'option *--fastq-list <filename>*.

- Pour traiter un échantillon particulier à partir du fichier CSV, utilisez l'option *--fastq-list-sample-id <SampleID>*. Seules les entrées du fichier fastq-list dont la valeur de RGSM correspond à la valeur de SampleID (identifiant d'échantillon) spécifiée sont traitées.
- Pour traiter tous les échantillons au cours d'une même analyse, sans égard à la valeur de RGSM, mettez à l'option *--fastq-list-all-samples* la valeur « true » (activée).

REMARQUE

Pour une analyse, seuls un fichier BAM de sortie et un fichier VCF de sortie sont produits, car il est attendu que tous les groupes de lectures entrés proviennent du même échantillon. Pour traiter plusieurs échantillons à partir d'une conversion en format BCL, lancez l'analyse secondaire de la plateforme DRAGEN plusieurs fois en utilisant différentes valeurs pour l'option *--fastq-list-sample-id*.

Aucune option ne permet de spécifier des groupes ou des sous-ensembles de valeurs RGSM pour mieux les filtrer, mais le fichier fastq-list peut être modifié pour atteindre le même objectif.

Voici un exemple de fichier CSV fastq-list présentant les colonnes requises :

```
RGID,RGSM,RGLB,Lane,Read1File,Read2File
   CACACTGA.1,RDSR181520,UnknownLibrary,1,/staging/RDSR181520_S1_L001_R1_
   001.fastq, /staging/RDSR181520_S1_L001_R2_001.fastq
   AGAACGGA.1,RDSR181521,UnknownLibrary,1,/staging/RDSR181521_S2_L001_R1_
   001.fastq, /staging/RDSR181521_S2_L001_R2_001.fastq
   TAAGTGCC.1,RDSR181522,UnknownLibrary,1,/staging/RDSR181522_S3_L001_R1_
   001.fastq, /staging/RDSR181522_S3_L001_R2_001.fastq
   AGACTGAG.1,RDSR181523,UnknownLibrary,1,/staging/RDSR181523_S4_L001_R1_
   001.fastq, /staging/RDSR181523_S4_L001_R2_001.fastq
```

Si vous utilisez l'option *--tumor-fastq-list* pour l'entrée en mode somatique, utilisez l'option *--tumor-fastq-list-sample-id <SampleID>* pour spécifier l'identifiant de l'échantillon pour la liste FASTQ correspondante, comme montré dans l'exemple suivant :

```
dragen -r <ref_dir> --tumor-fastq-list <csv_file> \
   --tumor-fastq-list-sample-id <Sample_ID> \
   --output-directory <out_dir> \
   --output-file-prefix <out_prefix> --fastq-list <csv_file_2> \
   --fastq-list-sample-id <Sample_ID_2>
```

Fichiers BAM d'entrée

Pour utiliser des fichiers BAM en entrée de la fonction de cartographie et d'alignement, mettez à l'option *--enable-map-align* la valeur « true » (activée). Si vous laissez cette option à « false » (désactivée), la valeur par défaut, vous pouvez utiliser le fichier BAM en entrée de la fonction de définition de variants.

Quand vous spécifiez un fichier BAM en entrée, la plateforme DRAGEN ignore tout renseignement sur l'alignement contenu dans le fichier en entrée et produit de nouveaux alignements pour toutes les lectures. Si le fichier en entrée contient des lectures appariées, il est important de spécifier que les données en entrée doivent être triées afin que les paires soient traitées ensemble. Dans les autres pipelines, il vous faudra trier de nouveau l'ensemble des données du fichier d'entrée par nom de lecture. La plateforme DRAGEN accélère considérablement cette opération en appariant les lectures en entrée, puis en les envoyant à la fonction de cartographie et d'alignement quand les paires ont été identifiées. Utilisez l'option *--pair-by-name* pour activer ou désactiver cette fonctionnalité (la valeur par défaut est « true » (activée)).

Spécifiez les lectures à paire de bases unique dans un fichier BAM d'entrée à l'aide des options *(-b)* et `--pair-by-name=false`, comme suit :

```
dragen -r <ref_dir> -b <bam> --output-directory <out_dir> \
   --output-file-prefix <out_prefix> --pair-by-name false
```

Spécifiez les lectures appariées dans un fichier BAM d'entrée à l'aide des options *(-b)* et `--pair-by-name=true`, comme suit :

```
dragen -r <ref_dir> -b <bam> --output-directory <out_dir> \
   --output-file-prefix <out_prefix> --pair-by-name true
```

Fichier CRAM d'entrée

Vous pouvez utiliser des fichiers CRAM en entrée pour les fonctions de cartographie et d'alignement et de définition de variants de la plateforme DRAGEN. Les fonctionnalités de la plateforme DRAGEN disponibles quand vous utilisez des fichiers CRAM en entrée sont les mêmes que quand vous utilisez des fichiers BAM.

L'option *--cram-reference* n'est plus nécessaire. Les fonctions de compression et de décompression de fichier CRAM utilisent la référence de la plateforme DRAGEN.

Les options suivantes sont utilisées pour fournir un fichier CRAM en entrée à la fonction de cartographie et d'alignement ou à la fonction de définition de variants :

- *--cram-input* – Le chemin d'accès et le nom du fichier CRAM.
- *--cram-input* – Un exemple d'utilisation de lectures appariées en entrée dans un fichier CRAM. Mettez également à l'option *--pair-by-name* la valeur « true » (activée).

  ```
  dragen -r <ref_dir> --cram-input <cram> --output-directory <out_dir> \
     --output-file-prefix <out_prefix> --pair-by-name true
  ```

Gestion de bases N

Une des techniques qu'utilise la plateforme DRAGEN pour optimiser la gestion des séquences peut mener au remplacement du score de qualité de base attribué aux définitions de bases N.

Quand vous utilisez les options *--fastq-n-quality* et *--fastq-offset*, les scores de qualité de base sont remplacés par une qualité de base fixe. Les valeurs par défaut pour ces deux options sont respectivement de « 2 » et « 33 ». Ces valeurs se combinent pour équivaloir à la qualité minimale d'Illumina de 35 (caractère ASCII « # »).

Noms de lecture pour les lectures appariées

Selon une convention commune, les noms de lecture peuvent contenir des suffixes (comme « /1 » ou « /2 ») pour indiquer de quelle lecture de la paire il s'agit. Pour les fichiers BAM d'entrée avec l'option *--pair-by-name*, la plateforme DRAGEN ignore ces suffixes pour trouver les noms de paire correspondants. Par défaut, la plateforme DRAGEN utilise le caractère de la barre oblique comme délimiteur pour ces suffixes et ignore « /1 » et « /2 » lorsqu'elle compare les noms. Par défaut, la plateforme DRAGEN supprime ces suffixes des noms de lecture initiaux.

La plateforme DRAGEN dispose des options suivantes pour contrôler l'utilisation des suffixes :

- Pour modifier le caractère de délimitation des suffixes, utilisez l'option *--pair-suffix-delimiter*. Les valeurs valides pour cette option comprennent les caractères de la barre oblique (/), du point (.) et du deux-points (:).
- Pour conserver le nom entier, y compris les suffixes, mettez à l'option *--strip-input-qname-suffixes* la valeur « false » (désactivée).
- Pour ajouter un nouvel ensemble de suffixes à tous les noms de lecture, mettez à l'option *--append-read-index-to-name* la valeur « true » (activée). Le caractère de délimitation est spécifié à l'aide de l'option *--pair-suffix-delimiter*. Par défaut, le caractère du délimiteur est la barre oblique, « /1 » et « /2 » sont donc ajoutés aux noms.

Fichiers d'annotations de gènes d'entrée

Quand vous traitez des données de séquençage de l'ARN, vous pouvez fournir un fichier d'annotations de gènes à l'aide de l'option *--annotation-file*. Fournir ce fichier améliore l'exactitude de l'étape de cartographie et d'alignement (consultez la section *Fichiers d'entrée* à la page 125). Le fichier doit être conforme à la spécification du format GFT/GFF et contenir les transcrits annotés qui correspondent au génome de référence utilisé pour la cartographie. Le format similaire GFF3 n'est actuellement pas pris en charge.

La plateforme DRAGEN peut accepter le fichier SJ.out.tab (consultez la section *Fichier SJ.out.tab* à la page 128) comme fichier d'annotations pour aider à guider la fonction d'alignement dans le mode de fonctionnement à deux passages.

Sexe de l'échantillon

Utilisez l'option *--sample-sex* pour spécifier le sexe de l'échantillon dans la ligne de commande.
Ce renseignement est transféré à toutes les fonctions de définition (définition de petits variants, de VNC, de variants structurels et de génotypage de répétitions). La fonction de définition de VNC a une fonctionnalité distincte d'inférence du sexe, mais la valeur spécifiée avec l'option *--sample-sex* a préséance sur le sexe inféré. L'exemple suivant montre comment spécifier le sexe de l'échantillon à l'aide de l'option *--sample-sex* :

```
--sample-sex MALE
--sample-sex FEMALE
```

Si l'option *--sample-sex* n'est pas spécifiée dans la ligne de commande, la fonction d'estimation de la ploïdie est exécutée par défaut pour déterminer le sexe.

Conservation ou retrait des étiquettes de BQSR

L'outil Picard Base Quality Score Recalibration (BQSR) produit des fichiers BAM de sortie qui contiennent des étiquettes BI et BD. L'outil BSQR calcule ces étiquettes par rapport à la séquence exacte d'une lecture. Si un fichier BAM contenant des étiquettes BI et BD est utilisé en entrée de la fonction de cartographie et d'alignement avec découpage de bases, les étiquettes BI et BD pourraient ne plus être valides.

Il est recommandé de retirer ces étiquettes lorsque vous utilisez des fichiers BAM en entrée. Pour retirer les étiquettes BI et BD, mettez à l'option *--preserve-bqsr-tags* la valeur « *false* » (désactivée). Si vous conservez les étiquettes, la plateforme DRAGEN vous avertit qu'il faut désactiver le découpage de bases.

Options relatives aux groupes de lectures

La plateforme DRAGEN suppose que toutes les lectures d'un fichier FASTQ font partie du même groupe de lectures. Le système DRAGEN crée un seul descripteur de groupe de lectures @RG dans l'en-tête du fichier BAM de sortie. Il est possible de spécifier les attributs standards de fichier BAM suivants :

Attribut	Argument	Description
ID	--RGID	L'identifiant du groupe de lectures. Si vous ajoutez des paramètres de groupe de lectures, l'argument RGID est obligatoire. Il s'agit de la valeur inscrite dans chaque enregistrement de fichier BAM de sortie.
LB	--RGLB	La librairie.
PL	--RGPL	La plateforme ou technologie utilisée pour produire les lectures. La spécification du format BAM permet les valeurs CAPILLARY, LS454, ILLUMINA, SOLID, HELICOS, IONTORRENT et PACBIO.
PU	--RGPU	L'unité de la plateforme, p. ex., flowcell-barcode.lane.
SM	--RGSM	L'échantillon.
CN	--RGCN	Le nom du centre de séquençage qui a produit les lectures.
DS	--RGDS	La description.
DT	--RGDT	La date où l'analyse a été effectuée.
PI	--RGPI	La taille moyenne prévue des inserts.

Si un de ces arguments est utilisé, le logiciel de la plateforme DRAGEN ajoute une étiquette RG à tous les enregistrements de sortie pour indiquer qu'ils font partie d'un groupe de lectures. L'exemple suivant montre une ligne de commande qui contient des paramètres de groupe de lectures :

```
dragen --RGID 1 --RGCN Broad --RGLB Solexa-135852 \
   --RGPL Illumina --RGPU 1 --RGSM NA12878 \
   -r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
   -1 SRA056922.fastq --output-directory /staging/tmp/\
   --output-file-prefix rg_example
```

Quand vous utilisez l'option *--fastq-list* pour mettre plusieurs groupes de lectures en entrée, les étiquettes BAM (et les autres) sont spécifiées pour chaque groupe de lectures en ajoutant des colonnes dans le fichier fastq_list.csv. Tous les en-têtes de colonne sont composés de quatre lettres majuscules qui commencent par « RG ». Les valeurs inscrites dans les colonnes pour chaque groupe de lectures sont ajoutées au fichier BAM de sortie avec une étiquette identique.

Options relatives à la licence

Pour supprimer le message d'état de la licence à la fin de l'analyse, utilisez l'option *--lic-no-print*. L'exemple suivant montre un message d'état de licence :

```
LICENSE_MSG| ======================================================
   LICENSE_MSG| License report
   LICENSE_MSG|   Genome status [ACxxxxxxxxxxx] : used 1263.9 Gbases
   since 2018-Feb-15 (1263886160894 bases, unlimited)
   LICENSE_MSG|   Genome  bases [ACxxxxxxxxxxx] : 202000000
   LICENSE_MSG|   Genome  bases [total]         : 202000000
```

Génération automatique de fichier MD5SUM pour les fichiers BAM et CRAM de sortie

Un fichier MD5SUM est généré automatiquement pour les fichiers BAM et CRAM de sortie. Le nom du fichier MD5SUM est composé du nom de fichier de sortie auquel et de l'extension .md5sum (p. ex. whole_genome_run_123.bam.md5sum). Le fichier MD5SUM est un fichier texte d'une ligne qui contient la valeur de md5sum du fichier de sortie. Celle valeur correspond exactement à la sortie de la commande md5sum de Linux.

La valeur de MD5SUM est calculée au moment de l'écriture du fichier de sortie. Il n'y a donc pas d'effet mesurable sur la performance (par rapport à la commande md5sum de Linux qui peut prendre plusieurs minutes à s'exécuter pour un fichier BAM de couverture 30x).

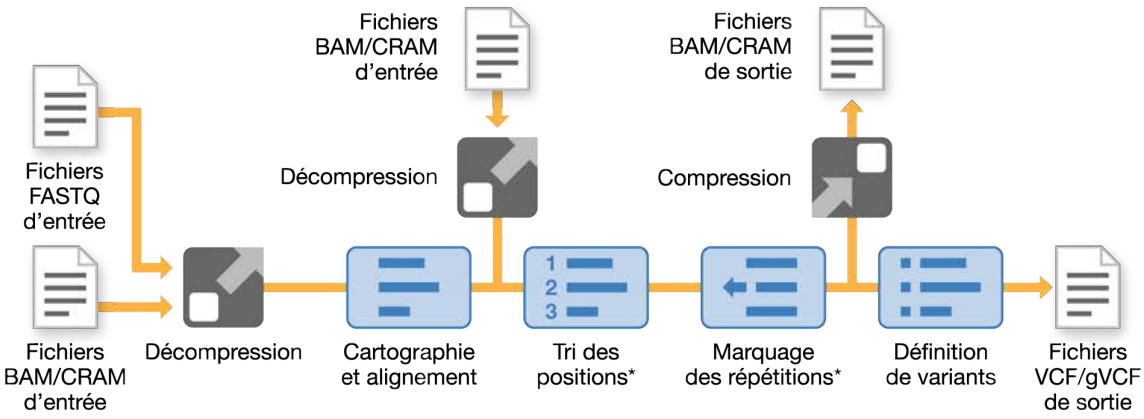
Fichiers de configuration

Les options de ligne de commande peuvent être stockées dans un fichier de configuration. L'emplacement du fichier de configuration par défaut est /opt/edico/config/dragen-user-defaults.cfg. Vous pouvez remplacer ce fichier en utilisant l'option *--config-file (-c)* pour spécifier un fichier différent. Le fichier de configuration utilisé pour une analyse donnée fournit les paramètres par défaut pour celle-ci. Tous ces paramètres peuvent être modifiés à l'aide des options de ligne de commande.

L'approche recommandée est d'utiliser le fichier dragen-user-defaults.cfg comme modèle pour créer des paramètres par défaut pour les différents cas d'utilisation. Copiez le fichier dragen-user-defaults.cfg, renommez la copie, puis modifiez le nouveau fichier pour un cas d'utilisation particulier. La pratique exemplaire est de mettre les options qui sont rarement modifiées dans le fichier de configuration et de spécifier les options qui varient d'une analyse à l'autre dans la ligne de commande.

Chapitre 3 Pipeline d'ADN de la plateforme DRAGEN

Figure 2 Pipeline d'ADN de la plateforme DRAGEN



* Facultatif

Cartographie de l'ADN

Option de densité de points d'ancrage

L'option *seed-density* contrôle combien de points d'ancrage primaires (qui normalement se chevauchent) de chaque lecture la fonction de cartographie recherche dans ses tables de hachage pour obtenir des correspondances exactes. La valeur de densité maximale de « 1,0 » génère un point d'ancrage commençant à n'importe quelle position dans la lecture, c'est-à-dire (L - K + 1) des points d'ancrage de base K à partir d'une lecture de base L.

La densité des points d'ancrage doit se trouver entre « 0,0 » et « 1,0 ». Dans la plateforme, une série de points d'ancrage équivalente à la densité demandée ou s'en rapprochant est sélectionnée. La série la plus clairsemée est d'un point d'ancrage par 32 positions, soit une valeur de densité de « 0,03125 ».

- Considérations relatives à l'exactitude – En général, la recherche de points d'ancrage dans des séries denses améliore l'exactitude de la cartographie. Toutefois, pour les lectures de longueur modeste (p. ex., 50 paires de bases et plus) et de faibles taux d'erreur de séquenceur, il y a peu d'amélioration à aller chercher au-delà de la densité de recherche de points d'ancrage par défaut de 50 %.
- Considérations relatives à la vitesse – Les séries denses de recherche de points d'ancrage ralentissent généralement le processus de cartographie, alors que les séries plus clairsemées ont tendance à l'accélérer. Toutefois, lorsque l'étape de cartographie des points d'ancrage peut être effectuée plus rapidement que l'étape d'alignement, alors une série plus clairsemée ne permet pas d'accélérer le processus de cartographie de beaucoup.

Relation avec l'intervalle de points d'ancrage de la référence

Fonctionnellement, un modèle de recherche de points d'ancrage plus dense ou plus épars a une incidence très semblable à un intervalle de points d'ancrage de référence plus long ou plus court (option de construction de la table de hachage *--ht-ref-seed-interval*). Générer 100 % des positions des points d'ancrage de référence et rechercher 50 % des positions de points d'ancrage de lecture a le même effet que

de générer 50 % des positions des points d'ancrage de référence et rechercher 100 % des positions de points d'ancrage de lecture. Dans les deux cas, la densité que l'on attend du nombre d'occurrences de point d'ancrage est 50 %.

Plus généralement, la densité que l'on attend du nombre d'occurrences de point d'ancrage est le produit de la densité des points d'ancrage de la référence (l'inverse de l'intervalle des points d'ancrage de la référence) et de la densité de recherche de points d'ancrage. Par exemple, si 50 % des points d'ancrage de la référence sont générés et que 33,3 % (1/3) des positions des points d'ancrage de la lecture sont recherchés, alors la densité que l'on attend des occurrences de point d'ancrage devrait être 16,7 % (1/6).

La plateforme DRAGEN ajuste automatiquement son modèle de recherche de points d'ancrage précis pour ne pas sauter systématiquement des positions de points d'ancrage générés à partir de la référence. Par exemple, la fonction de cartographie ne recherche pas les points d'ancrage qui ne correspondent qu'à des positions impaires dans la référence lorsque seules des positions paires sont générées dans la table de hachage, même si l'intervalle des points d'ancrage de la référence est « 2 » et que la densité des points d'ancrage est « 0,5 ».

Option d'orientation de la cartographie

L'option *--Mapper.map-orientations* est utilisée pour effectuer la cartographie de lectures aux fins d'analyse de la méthylation de l'ADN lors du séquençage au bisulfite. Elle est définie automatiquement selon la valeur de l'option *--methylation-protocol*.

L'option *--Mapper.map-orientations* permet de restreindre l'orientation cartographique des lectures dans le génome de référence dans l'ordre normal seulement ou en complément inverse seulement. Les valeurs valides pour l'option *--map-orientations* sont les suivantes :

- 0 – La valeur par défaut
- 1 – La cartographie dans l'ordre normal seulement
- 2 – La cartographie en ordre de complément inverse seulement

Si les orientations de la cartographie sont restreintes et que des lectures appariées sont utilisées, l'orientation attendue des paires ne peut être que l'orientation par défaut et non dans l'ordre normal ou en complément inverse seulement.

Options de modification des points d'ancrage

Bien que la plateforme DRAGEN cartographie d'abord les lectures en cherchant les correspondances exactes de la référence à des points d'ancrage courts, elle peut aussi cartographier des points d'ancrage qui diffèrent de la référence d'un nucléotide en cherchant aussi des points d'ancrage modifiés par polymorphisme mononucléotidique. La modification de points d'ancrage n'est habituellement pas nécessaire avec des lectures longues (100 paires de bases et plus), car elles ont une probabilité élevée de contenir au moins une correspondance exacte de point d'ancrage. Cela est particulièrement vrai lorsque des lectures appariées sont utilisées, car la correspondance d'un point d'ancrage à l'une des deux lectures permet de bien aligner la paire. Mais la modification de points d'ancrage peut, par exemple, être utile pour améliorer l'exactitude de la cartographie pour les lectures courtes à paire de bases unique. Toutefois, la cartographie prendra plus de temps à exécuter. Les options suivantes contrôlent la modification des points d'ancrage :

Tableau 1 Options de modification des points d'ancrage

Nom de l'option de ligne de commande	Nom de l'option du fichier de configuration
--Mappper.seed-density	seed-density
--Mapper.edit-mode	edit-mode

Nom de l'option de ligne de commande	Nom de l'option du fichier de configuration
--Mapper.edit-seed-num	edit-seed-num
--Mapper.edit-read-len	edit-read-len
--Mapper.edit-chain-limit	edit-chain-limit

Options edit-mode et edit-chain-limit

Les options edit-mode et edit-chain-limit contrôlent l'utilisation de la modification des points d'ancrage. Les quatre valeurs suivantes peuvent être utilisées pour l'option edit-mode :

Mode	Description
0	Aucune modification (valeur par défaut)
1	Test de longueur de la chaîne
2	Test de longueur de la chaîne appariée
3	Modification complète des points d'ancrage

Le mode de modification 0 nécessite que tous les points d'ancrage correspondent exactement. Le mode 3 est le plus coûteux, car chaque point d'ancrage qui ne correspond pas exactement à la référence est modifié. Les modes 1 et 2 emploient des méthodes heuristiques pour rechercher des points d'ancrage modifiés seulement pour les lectures qui ont les meilleures probabilités d'être récupérées aux fins d'exactitude de la cartographie.

La principale méthode heuristique des modes de modification 1 et 2 est un test de longueur de la chaîne de points d'ancrage. Les points d'ancrage exacts sont cartographiés à la référence au premier passage sur une lecture donnée, puis les points d'ancrage correspondants sont regroupés sous forme de chaînes de points d'ancrage aux alignements semblables. Si la chaîne de points d'ancrage la plus longue (dans la lecture) excède un seuil edit-chain-limit, la lecture est jugée comme n'ayant pas besoin de modification de points d'ancrage, car elle se trouve déjà en position prometteuse pour la cartographie.

Le mode de modification 1 déclenche la modification pour une lecture donnée à l'aide du test de longueur de la chaîne de points d'ancrage. Si aucune chaîne n'excède le seuil *edit-chain-limit* (y compris s'il n'y a aucune correspondance exacte de points d'ancrage), alors un second passage de cartographie est tenté avec des points d'ancrage modifiés. Le mode de modification 2 optimise encore plus la méthode heuristique pour les lectures appariées. Si l'une des deux lectures de la paire a une chaîne de points d'ancrage plus longue que le seuil *edit-chain-limit*, alors la modification des points d'ancrage est désactivée pour cette paire, car une analyse de récupération permettra probablement de récupérer l'alignement de la lecture à partir des correspondances de points d'ancrage tirés d'une lecture. Le mode de modification 2 est le même que le mode 1 pour les lectures à paire de bases unique.

Options edit-seed-num et edit-read-len

Pour les modes 1 et 2, lorsque la méthode heuristique déclenche la modification, ces options contrôlent combien de positions de points d'ancrage sont modifiées lors du deuxième passage sur la lecture. Bien que la cartographie exacte des points d'ancrage puisse utiliser une série dense de points d'ancrage se chevauchant, comme des points d'ancrage commençant à 50 % ou à 100 % des positions de la lecture, une grande partie de la valeur de modification des points d'ancrage peut être obtenue en modifiant une série beaucoup plus clairsemée, ou même une série de points ne se chevauchant pas.

En général, si une application d'utilisateur peut se permettre de passer une plus grande partie du temps de cartographie à la modification des points d'ancrage, l'exactitude de la cartographie peut être améliorée (pour le même temps) en modifiant les points d'ancrage des séries clairsemées pour un grand nombre de lectures plutôt qu'en modifiant des points d'ancrage en séries denses pour un petit nombre de lectures.

Lorsque la modification est déclenchée, ces deux options nécessitent les positions de modification des points d'ancrage *edit-seed-num* qui sont distribuées uniformément parmi les premières bases *edit-read-len* de la lecture. Par exemple, avec des points d'ancrage de 21 bases, l'option *edit-seed-num* à « 6 » et l'option *edit-read-len* à « 100 », les points d'ancrage modifiés peuvent être décalés {0, 16, 32, 48, 64, 80} par rapport à la fin du 5', les points d'ancrage chevauchant 5 bases. Puisque les technologies de séquençage produisent souvent des bases de qualité près du début (5') de chaque lecture, cela permet de concentrer la modification de points d'ancrage là où elle a le plus de chances de réussir. Lorsqu'une lecture est plus courte que la valeur de l'option *edit-read-len*, un plus petit nombre de points d'ancrage est modifié.

La modification des points d'ancrage coûte plus cher lorsque l'intervalle de points d'ancrage de référence (option de construction de la table de hachage *--ht-ref-seed-interval*) est supérieur à « 1 ». Pour les modes 1 et 2, des positions de modification de points d'ancrage supplémentaires sont automatiquement générées afin de ne pas manquer les positions de points d'ancrage de la référence. Pour le mode 3, la durée et les coûts peuvent augmenter de manière importante, car les points d'ancrage recherchés qui correspondent à des positions vides de la référence sautent généralement ces positions, ce qui déclenche la modification.

Alignement de l'ADN

Paramètres du score d'alignement Smith-Waterman

La première étape de la cartographie est de générer des points d'ancrage à partir de la lecture et de rechercher des correspondances exactes dans le génome de référence. Ces résultats sont ensuite raffinés en exécutant des alignements Smith-Waterman aux emplacements ayant la densité de correspondances de points d'ancrage la plus élevée. Cet algorithme bien documenté compare chaque position de la lecture avec les positions candidates de la référence. Ces comparaisons correspondent à une matrice d'alignements potentiels entre la lecture et la référence. Pour chacune de ces positions d'alignement candidates, l'algorithme Smith-Waterman génère des scores qui sont utilisés pour évaluer si le meilleur alignement passant par cette cellule matrice le fait grâce à une correspondance ou un mésappariement de nucléotides (mouvement diagonal), une délétion (mouvement horizontal) ou une insertion (mouvement vertical). Une correspondance entre la lecture et la référence bonifie le score, alors qu'une non-correspondance ou un indel lui impose une pénalité. Le chemin vers le score global le plus élevé, dans la matrice, est celui de l'alignement choisi.

Les valeurs choisies pour les scores de cet algorithme indiquent comment équilibrer, pour un alignement ayant plusieurs possibilités d'interaction, la possibilité d'un indel en comparaison avec au moins un polymorphisme mononucléotidique, ou la préférence d'un alignement sans découpage. Les valeurs de notation par défaut de la plateforme DRAGEN permettent d'aligner des lectures de longueurs modérées avec un génome humain de référence entier aux fins d'utilisation en définition de variants. Toutefois, n'importe quel ensemble de paramètres de notation de Smith-Waterman représente un modèle imprécis de mutation génomique et d'erreurs de séquençage et des valeurs de notation d'alignement autrement paramétrées peuvent se révéler plus appropriées pour certaines utilisations.

Les options d'alignement suivantes contrôlent l'alignement de Smith-Waterman :

Nom de l'option de ligne de commande	Nom de l'option du fichier de configuration
--Aligner.global	global
--Aligner.match-score	match-score
--Aligner.match-n-score	match-n-score
--Aligner.mismatch-pen	mismatch-pen
--Aligner.gap-open-pen	gap-open-pen
--Aligner.gap-ext-pen	gap-ext-pen
--Aligner.unclip-score	unclip-score
--Aligner.no-unclip-score	no-unclip-score
--Aligner.aln-min-score	aln-min-score

- *global*

  L'option *global* (la valeur peut être « 0 » ou « 1 ») contrôle si l'alignement doit absolument se faire sur toute la longueur dans la lecture. Lorsque la valeur est « 1 », les alignements sont toujours faits sur toute la longueur, comme dans l'algorithme d'alignement global de Needleman-Wunsch (bien qu'ils ne soient pas de bout en bout dans la référence), et les scores d'alignement peuvent être positifs ou négatifs. Lorsque la valeur est « 0 », les alignements peuvent être coupés à au moins une des deux extrémités de la lecture, comme dans l'algorithme d'alignement local de Smith-Waterman, et les scores d'alignements ne peuvent pas être négatifs.

  En général, une valeur de l'option *global* de « 0 » est préférable pour les lectures longues pour que les segments importants de la lecture qui suivent une coupure quelconque (notamment gros indel, variant structurel, lecture chimérique) puissent être coupés sans pénaliser gravement le score d'alignement. Mettre la valeur de l'option *global* de « 1 » peut ne pas avoir l'effet escompté sur des lectures longues, car des insertions effectuées à l'extrémité ou presque d'une lecture peuvent être perçues comme du pseudodécoupage. Aussi, si la valeur de l'option *global* est de « 0 », les alignements multiples (chimériques) peuvent être rapportés lorsque de grandes portions d'une lecture correspondent à des positions de référence très séparées.

  Mettre la valeur de l'option *global* à « 1 » est parfois préférable pour les lectures courtes, qui ont peu de probabilité de chevaucher des ruptures structurelles, qui sont incapables de prendre en charge les alignements chimériques et qui sont soupçonnées de mauvaise cartographie lorsqu'elles ne peuvent pas bien être alignées sur toute la longueur.

  Pensez à utiliser l'option *unclip-score*, ou à en augmenter la valeur, plutôt que de mettre à l'option *global* la valeur « 1 » afin de marquer une légère préférence pour les alignements non coupés.

- *match-score*

  L'option *match-score* est le score du nucléotide d'une lecture qui correspond à un nucléotide de référence (A, C, G ou T). Sa valeur est un nombre entier signé entre « 0 » et « 15 ». La valeur de l'option *match_score* de « 0 » peut seulement être utilisée qu'avec la valeur de l'option *global* de « 1 ». Un score de correspondance plus élevé entraîne des alignements plus longs et un nombre moins élevé de longues insertions.

- *match-2-score*

  L'option match-2-score est le score du nucléotide d'une lecture qui correspond à un code IUPAC-IUB à deux bases dans la référence (K, M, R, S, W ou Y). La valeur de cette option est un nombre entier signé entre « -16 » et « 15 ».

- *match-3-score*

  L'option *match-3-score* est le score du nucléotide d'une lecture qui correspond à un code IUPAC-IUB à trois bases dans la référence (B, D, H ou V). La valeur de cette option est un nombre entier signé entre « -16 » et « 15 ».

- *match-n-score*

  L'option *match-n-score* est le score du nucléotide d'une lecture qui correspond à un code « N » dans la référence. La valeur de cette option est un nombre entier signé entre « -16 » et « 15 ».

- *mismatch-pen*

  L'option *mismatch-pen* correspond à la pénalité (score négatif) pour le nucléotide d'une lecture mésappariée à un nucléotide de la référence ou à un code IUPAC-IUB (sauf « N », qui ne peut être mésapparié). La valeur de cette option est un nombre entier non signé entre « 0 » et « 63 ». Une pénalité de mésappariement plus élevée donnera un alignement présentant davantage d'insertions, de délétions et de découpages afin d'éviter les polymorphismes mononucléotidiques.

- *gap-open-pen*

  L'option *gap-open-pen* correspond à la pénalité (score négatif) pour l'intégration d'un écart (c.-à-d. une insertion ou une délétion). Cette valeur n'est valide que pour un écart de 0 base. Elle est toujours ajoutée à la longueur de l'écart gap-ext-pen. La valeur de cette option est un nombre entier non signé entre « 0 » et « 127 ». Une pénalité plus élevée d'intégration d'un écart entraîne moins d'insertions et de délétions, toutes longueurs confondues, dans les alignements CIGAR. Le découpage ou l'alignement par polymorphisme mononucléotidique sont plutôt utilisés.

- *gap-ext-pen*

  L'option *gap-ext-pen* correspond à la pénalité (score négatif) pour l'extension d'un écart (c.-à-d. une insertion ou une délétion) par une base. La valeur de cette option est un nombre entier non signé entre « 0 » et « 15 ». Une pénalité plus élevée d'extension d'un écart entraîne moins de longues insertions et délétions dans les alignements CIGAR. Les indels courts, le découpage ou l'alignement par polymorphisme mononucléotidique sont plutôt utilisés.

- *unclip-score*

  L'option *unclip-score* est la bonification du score pour un alignement qui atteint le début ou la fin d'une lecture. Un alignement sur toute la longueur reçoit deux fois cette bonification. La valeur de cette option est un nombre entier non signé entre « 0 » et « 127 ». Une bonification de non-coupage plus élevée signifie que l'alignement atteint le début ou la fin d'une lecture plus souvent, alors que le résultat peut être atteint sans utiliser trop de polymorphismes mononucléotidiques ou d'indels.

  Un score unclip-score différent de « 0 » est utile lorsque la valeur de *global* est de « 0 » pour marquer une légère préférence pour les alignements non coupés. La bonification non coupée a peut d'incidence sur les alignements la valeur de *global* est de « 1 », car les alignements faits sur toute la longueur sont forcés de toute façon (bien que 2 x unclip-score s'ajoute à n'importe quel score d'alignement à moins que la valeur de *no-unclip-score* soit de « 1 »). L'utilisation de la valeur par défaut de l'option unclip-score est recommandée lorsque la valeur de *global* est « 1 , car certaines méthodes heuristiques internes prennent en considération comment les alignements locaux ont été coupés.

  Sachez que, en particulier avec des lectures longues, spécifier un score unclip-score beaucoup plus élevé que la valeur de gap-open-pen peut causer des effets indésirables, soit l'utilisation d'insertions à une extrémité ou près d'une extrémité d'une lecture en guise de pseudodécoupage, surtout si la valeur de global est « 1 ».

- *no-unclip-score*

  La valeur de l'option *no-unclip-score* peut être « 0 » ou « 1 ». La valeur par défaut est « 1 ». Lorsque la valeur de l'option *no-unclip-score* est « 1 », toute bonification de non-coupage (unclip-score) contribuant à un alignement est retirée du score de l'alignement avant de poursuivre son traitement, comme la comparaison avec *aln-min-score*, la comparaison avec d'autres scores d'alignements et le signalement des étiquettes AS ou XS. Toutefois, la bonification de non-coupage a tout de même une incidence sur l'alignement au meilleur score trouvé par l'algorithme de Smith-Waterman par rapport à un segment de référence donné, créant un biais envers les alignements non coupés.

  Lorsque la valeur de l'option unclip-score est supérieure à « 0 » et entraîne l'extension d'un alignement local de Smith-Waterman à au moins l'une des extrémités de la lecture, le score d'alignement reste le même ou augmente si la valeur *no-unclip-score* est « 0 », ou il reste le même ou diminue si la valeur de *no-unclip-score* est « 1 ».

  La valeur d'option par défaut, *no-unclip-score* de « 1 », est recommandée lorsque la valeur de *global* est de « 1 », car chaque alignement est fait sur toute la longueur et il n'est pas nécessaire d'ajouter la même bonification à chaque alignement.

  En changeant la valeur de l'option *no-unclip-score*, pensez à ajuster la valeur de *aln-min-score*. Lorsque la valeur de *no-unclip-score* est « 0 », les bonifications de non-coupage sont comprises dans les scores d'alignement en comparaison avec le score minimal *aln-min-score*, de manière à ce que le sous-ensemble d'alignements filtrés par l'option *aln-min-score* peut changer de manière importante grâce à l'option *no-unclip-score*.

- *aln-min-score*

  L'option *aln-min-score* spécifie un score d'alignement minimal acceptable. Tous les alignements dont le score est inférieur à sa valeur sont rejetés. L'augmentation ou la diminution de la valeur de *aln-min-score* permet de réduire ou d'augmenter le pourcentage des lectures cartographiées. La valeur de cette option est un nombre entier signé (les scores d'alignement négatifs sont possibles lorsque la valeur de *global* est « 0 »).

  La valeur de *aln-min-score* a également une incidence sur les estimations du score MAPQ. La principale contribution au calcul du score MAPQ est la différence entre le premier et le deuxième meilleurs scores d'alignement et l'option *aln-min-score* sert de score d'alignement sous-optimal si le deuxième meilleur score est inférieur à la valeur de l'option. Par conséquent, l'augmentation de la valeur de *aln-min-score* peut diminuer le score MAPQ rapporté pour certains alignements à faibles scores.

Options relatives à l'appariement

La plateforme DRAGEN peut traiter des données appariées, soit à partir d'une paire de fichiers FASTQ, soit à partir d'un seul fichier FASTQ imbriqué. Elle cartographie les deux extrémités séparément, puis sélectionne un ensemble d'alignements qui présente la meilleure probabilité d'être une paire dans l'orientation attendue et ayant à peu près la taille d'inserts attendue. Les alignements des deux extrémités sont évalués pour la qualité de l'appariement, avec des pénalités plus élevées pour les tailles d'inserts qui ne correspondent pas à la taille attendue. Les options suivantes contrôlent le traitement des données appariées :

- *Réorientation*

  L'option *pe-orientation* spécifie l'orientation attendue des paires. Seules les paires respectant cette orientation peuvent être marquées comme paires valides. Les valeurs possibles sont les suivantes :

  - 0 – Ordre normal et de complément inverse (valeur par défaut)
  - 1 – Ordre de complément inverse
  - 2 – Ordre normal seulement

- *unpaired-pen*

  Les meilleures positions cartographiques sont déterminées conjointement pour chaque paire de lectures appariées, selon le score d'appariement le plus élevé, en considérant les différentes combinaisons d'alignements pour chaque lecture du couple. Un score d'appariement correspond à la somme des scores de deux alignements moins une pénalité d'appariement, qui estime la probabilité que la longueur de l'insert soit plus éloignée de celle de l'insert moyen que cette paire.

  L'option *unpaired-pen* spécifie la pénalité des scores d'appariement lorsque les deux alignements ne sont pas dans une position ou une orientation valide. Cette option sert aussi de pénalité d'appariement maximale pour les alignements bien appariés dont les longueurs d'inserts sont extrêmes.

  L'option *unpaired-pen* est spécifiée à l'échelle Phred, selon son incidence potentielle sur le score MAPQ. À l'interne, elle est adaptée à l'espace du score d'alignement selon les paramètres d'attribution de scores de Smith-Waterman.

- *pe-max-penalty*

  L'option *pe-max-penalty* limite l'augmentation de l'estimation du score MAPQ pour une lecture lorsque l'autre lecture du couple se trouve à proximité. Un alignement apparié ne reçoit jamais un score MAPQ supérieur au score MAPQ qui lui serait attribué pour une cartographie à paire de bases unique, plus cette valeur. Par défaut, la valeur de l'option *pe-max-penalty* à celle de l'option *mapq-max* qui est « 255 », ce qui désactive cette limite.

  La principale différence entre l'option *unpaired-pen* et l'option *pe-max-penalty* est que *unpaired-pen* a une incidence sur le calcul des scores des paires, donc la sélection des alignements, alors que *pe-max-penalty* n'a d'incidence que sur les scores MAPQ rapportés pour les alignements appariés.

Détection de la taille moyenne des inserts

Pour traiter des données appariées, la plateforme DRAGEN doit faire un choix entre les alignements de la meilleure qualité possible aux deux extrémités afin de sélectionner les paires probables. Pour faire ce choix, la plateforme utilise un modèle statistique gaussien pour évaluer les probabilités que deux alignements constituent une paire. Ce modèle est basé sur l'intuition que la préparation d'une librairie tend à créer des fragments de tailles semblables, produisant ainsi des paires dont les longueurs d'inserts se regroupent bien autour d'une longueur moyenne d'insert.

Si vous connaissez les statistiques de votre préparation de librairie pour un fichier d'entrée (et que le fichier consiste en un seul groupe de lectures), vous pouvez spécifier les caractéristiques de la distribution des longueurs d'inserts : longueur moyenne, écart-type et pour les trois quartiles. Ces caractéristiques peuvent être spécifiées à l'aide des options *Aligner.pe-stat-mean-insert*, *Alinger.pe-stat-stddev-insert*, *Aligner.pe-stat-quartiles-insert* et *Aligner.pe-stat-mean-read-len*. Il est toutefois préférable de laisser la plateforme DRAGEN les détecter automatiquement.

Pour activer l'échantillonnage automatique de la distribution des longueurs d'inserts, mettez à l'option *--enable-sampling* la valeur « true » (activée). Lorsque le logiciel lance l'exécution, il traite un échantillon composé de jusqu'à 100 000 paires grâce à la fonction d'alignement, calcule la distribution, puis utilise les résultats statistiques pour évaluer chacune des paires de l'ensemble d'entrée.

Le logiciel hôte de la plateforme DRAGEN rapporte les statistiques dans son journal stdout comme suit :

```
Final paired-end statistics detected for read group 0, based on 79935
   high quality pairs for FR orientation
   Quartiles (25 50 75) = 398 410 421
   Mean = 410.151
   Standard deviation = 14.6773
```

```
    Boundaries for mean and standard deviation: low = 352, high = 467
    Boundaries for proper pairs: low = 329, high = 490
NOTE: DRAGEN's insert estimates include corrections for clipping (so they are
no identical to TLEN)
```

La distribution des longueurs d'inserts pour chaque échantillon est inscrite dans le fichier fragment_length_hist.csv. Chaque échantillon commence avec les lignes suivantes :

```
#Sample: sample name
FragmentLength,Count
```

Ces lignes sont suivies par l'histogramme.

Lorsque le nombre de paires est très petit, il n'est pas possible de caractériser la distribution avec confiance. Dans ce cas, la plateforme DRAGEN applique des statistiques par défaut qui indiquent une distribution d'inserts très large et tend à considérer des paires d'alignements comme étant de vraies paires, même si ces alignements se trouvent à des dizaines de milliers de bases l'un de l'autre. Dans cette situation, la plateforme produit un message, comme suit :

```
WARNING: Less than 28 high quality pairs found - standard deviation is
    calculated from the small samples formula
```

La formule pour petits échantillons calcule l'écart-type comme suit :

```
if samples < 3 then
     standard deviation = 10000
else if samples < 28 then
    standard deviation = 25 * (standard deviation + 1)/(samples - 2)
end if
if standard deviation < 12 then
     standard deviation = 12
end if
```

Le modèle par défaut est « standard deviation = 10000 » (écart-type = 10 000). Si les 10 000 premières lectures ne sont pas cartographiées ou que toutes les paires sont mauvaises, l'écart-type a la valeur « 10000 » et la moyenne et les quartiles à « 0 ». Remarque : La valeur minimale pour l'écart-type est « 12 », indépendamment du nombre d'échantillons.

Pour des données de séquençage de l'ARN, la distribution des tailles d'inserts n'est pas normale puisque des paires contiennent des introns. Le logiciel DRAGEN estime la distribution à l'aide d'une fonction d'estimation à noyaux de la densité pour offrir une longue queue aux échantillons. Cette estimation améliore l'exactitude de la moyenne et de l'écart-type pour les données de séquençage de l'ARN et le bon appariement.

Analyses de récupération

Pour les lectures appariées, lorsqu'une occurrence de point d'ancrage pour une des deux lectures d'une paire mais pas l'autre, les analyses de récupération recherchent des alignements manquants pour l'autre lecture, dans une portée de récupération de la longueur de la moyenne des inserts. Normalement, le logiciel hôte de la plateforme DRAGEN a une portée de récupération de 2,5 écarts-types de la distribution des inserts empiriques. Mais, dans les cas où l'écart-type des inserts est plus grand que la longueur de la lecture, la portée de récupération est restreinte de manière à limiter les ralentissements durant la cartographie.

Un message d'avertissement s'affiche alors, comme suit :

```
Rescue radius = 220
     Effective rescue sigmas = 0.5
          WARNING: Default rescue sigmas value of 2.5 was overridden by host software!
          The user may wish to set rescue sigmas value explicitly with --Aligner.rescue-sigmas
```

Bien que l'utilisateur puisse ignorer cet avertissement, ou spécifier une portée de récupération intermédiaire afin de ne pas ralentir la cartographie, il est recommandé d'utiliser 2,5 facteurs sigma pour que la portée de récupération conserve sa sensibilité durant la cartographie. Pour désactiver l'analyse de récupération, mettez à l'option *max-rescues* la valeur « 0 ».

Options relatives à la sortie

La plateforme DRAGEN peut effectuer le suivi de quatre alignements indépendants pour chaque lecture. Ces alignements peuvent être chimériques, ce qui signifie qu'ils peuvent être cartographiés à différentes régions de la lecture, ou des cartographies sous-optimales de la lecture à différentes régions de l'alignement.

Lorsque la plateforme DRAGEN effectue le suivi des quatre meilleurs alignements pour une lecture, elle priorise les alignements chimériques (supplémentaires) par rapport aux alignements sous-optimaux (secondaires). En d'autres termes, elle effectue le suivi d'autant d'alignements chimériques que possible en respectant sa limite de quatre. S'il y a moins de quatre alignements chimériques, alors les alignements restants seront sous-optimaux.

Vous pouvez utilisez les options de configuration suivantes pour contrôler combien d'alignements de chaque type vous voulez inclure dans le fichier de sortie de la plateforme DRAGEN.

- *mapq-max*

  L'option *mapq-max* spécifie un plafond pour l'estimation du score MAPQ qui peut être rapporté pour un alignement, entre 0 et 255. Si le score MAPQ calculé est plus élevé, c'est plutôt cette valeur qui est rapportée. La valeur par défaut est « 60 ».

- *supp-aligns*, *sec-aligns*

  Les options *supp-aligns* et *sec-aligns* restreignent, respectivement, le nombre maximal d'alignements supplémentaires (c.-à-d. les alignements chimériques et les alignements marqués SAM FLAG « 0x800 ») et d'alignements secondaires (c.-à-d. les alignements sous-optimaux et les alignements marqués SAM FLAG « 0x100 ») qui peuvent être rapportés pour chaque lecture.

  Un maximum de 31 alignements est rapporté pour un nombre total de lectures, qu'ils soient primaires, supplémentaires ou secondaires. C'est pourquoi les valeurs des options *supp-aligns* et *sec-aligns* vont de « 0 » à « 30 ». Les alignements supplémentaires ont la priorité sur les alignements secondaires aux fins de suivi et de rapport.

  Des valeurs élevées ont une incidence sur la vitesse pour ces deux options; il est recommandé de ne les augmenter qu'au besoin.

- *sec-phred-delta*

  L'option *sec-phred-delta* contrôle quels alignements secondaires sont produits selon le score d'alignement relatif à l'alignement primaire rapporté. Seuls les alignements secondaires dont les probabilités se situent dans la même échelle de valeurs Phred que l'alignement primaire sont signalés.

- *sec-aligns-hard*

  L'option *sec-aligns-hard* permet de supprimer les résultats relatifs à tous les alignements secondaires si leur nombre est plus élevé que le nombre qui peut être produit. Mettre à l'option *sec-aligns-hard* la valeur « 1 » permet de forcer les lectures à être non cartographiées lorsque les alignements secondaires ne produisent pas tous des résultats.

- *supp-as-sec*

  Lorsque la valeur de l'option *supp-as-sec* est mise à « 1 », les alignements supplémentaires (chimériques) sont rapportés avec la marque SAM FLAG « 0x100 » au lieu de « 0x800 ». La valeur par défaut est « 0 ». L'option *supp-as-sec* offre une compatibilité avec des outils qui ne prennent pas en charge les alignements FLAG « 0x800 ».

- *hard-clips*

  L'option *hard-clips* est utilisée sous forme de champ de 3 bits, avec des valeurs allant de « 0 » à « 7 ». Les bits spécifient les alignements comme suit :

  - Bit 0 – alignements primaires
  - Bit 1 – alignements supplémentaires
  - Bit 2 – alignements secondaires

  Chaque bit détermine quels types d'alignements locaux sont rapportés avec une base coupée absente (1) ou présente (0) dans l'alignement. La valeur par défaut est « 6 », ce qui signifie que les alignements primaires utilisent une base coupée présente dans l'alignement, alors que les alignements secondaires utilisent une base absente de l'alignement.

Cartographie sensible à la valeur du champ ALT

Le génome humain de référence GRCh38 contient beaucoup plus d'haplotypes alternatifs (contigs ALT) que ses anciennes versions. En général, l'inclusion des contigs ALT dans la référence cartographique améliore la spécificité de la cartographie et de la définition de variants, car les mauvais alignements sont éliminés des lectures qui correspondent à un contig ALT mais qui se comparent mal à l'assemblage primaire. Toutefois, la cartographie à l'aide des contigs ALT du génome GRCh38 sans traitement particulier peut nuire de façon importante à la sensibilité de la définition de variants dans les régions correspondantes, car plusieurs lectures peuvent s'aligner à un contig ALT et à la position correspondante dans l'assemblage primaire.
La cartographie sensible à la valeur du champ ALT de la plateforme DRAGEN élimine ce problème et réussit à améliorer à la fois la sensibilité et la spécificité des contigs ALT.

La cartographie sensible à la valeur du champ ALT nécessite des tables de hachage construites avec des alignements de fichiers LiftOver ALT spécifiés (consultez la section *Tables de hachage sensibles à la valeur du champ ALT* à la page 145). Si une table de hachage construite avec des alignements de fichiers LiftOver est fournie, la plateforme DRAGEN exécute automatiquement la cartographie sensible à la valeur du champ ALT. Mettre la valeur « false » (désactivée) à l'option *--alt-aware* permet de désactiver la cartographie sensible à la valeur du champ ALT avec un fichier de référence LiftOver.

La plateforme DRAGEN nécessite des tables de hachage sensibles à la valeur du champ ALT pour toute référence hg19 ou GRCh38 dans laquelle des contigs ALT ont été détectés. Pour désactiver cette exigence sur la plateforme DRAGEN, mettez à l'option *--ht-alt-aware-validate* la valeur « false » (désactivée).

Lorsque la cartographie sensible à la valeur du champ ALT est activée, les fonctions de cartographie et d'alignement sont sensibles à la relation entre les positions des contigs ALT et les positions correspondantes dans l'assemblage primaire. Les correspondances des points d'ancrage dans les contigs ALT sont utilisées pour obtenir des alignements dans l'assemblage primaire correspondant, même si ce dernier présente un faible score. Des groupes LiftOver sont formés, chacun contenant un alignement candidat à l'assemblage primaire et au moins zéro alignement ALT candidat visant le même emplacement. Chaque groupe LiftOver se voit attribuer un score selon ses alignements les mieux appariés, en prenant en comptes les alignements correctement appariés. Le groupe LiftOver gagnant fournit le représentant de son assemblage primaire comme alignement primaire de sortie. Le score MAPQ est calculé selon la différence entre le score de ce groupe et celui du groupe LiftOver arrivant en deuxième position. L'émission d'alignements primaires dans l'assemblage primaire préserve la couverture d'alignement de référence et facilite la définition de variants.

Si l'option *--Aligner.en-alt-hap-aln* est spécifiée avec une valeur de « 1 » et que l'option *--Aligner.supp-aligns* a une valeur supérieure à « 0 », les alignements d'haplotypes alternatifs correspondants peuvent aussi être produits dans le fichier de sortie. Ils seront marqués comme des alignements supplémentaires.

Voici un comparatif des autres options de traitement d'haplotypes alternatifs.

- Cartographie sans contigs ALT dans la référence :
  - La définition de variants donne de faux positifs lorsque des lectures appariées à un haplotype alternatif sont mal alignées ailleurs.
  - Une faible sensibilité de cartographie et de définition de variants donne des lectures appariées à un contig ALT, mais qui diffèrent fortement de l'assemblage primaire.
- Cartographie avec contigs ALT mais pas de sensibilité à la valeur du champ ALT :
  - Les définitions de variants donnant de faux positifs à cause de lectures mal alignées appariées à des contigs ALT sont éliminées.
  - Une couverture d'alignement nulle ou faible dans les régions de l'assemblage primaire couvertes par des haplotypes alternatifs, en raison de lectures cartographiées avec des contigs ALT.
  - Un score MAPQ nul ou faible dans les régions couvertes par des haplotypes alternatifs lorsqu'ils sont semblables ou identiques à l'assemblage primaire.
  - La sensibilité de la définition de variants est réduite de beaucoup dans les régions couvertes par des haplotypes alternatifs.
- Cartographie avec contigs ALT et sensibilité à la valeur du champ ALT :
  - Les définitions de variants donnant de faux positifs à cause de lectures mal alignées appariées à des contigs ALT sont éliminées.
  - La couverture des alignements de référence dans les régions couvertes par des haplotypes alternatifs, car les alignements primaires sont effectués avec l'assemblage primaire.
  - Des scores MAPQ de référence sont attribués parce que les candidats à l'alignement dans un groupe LiftOver ne sont pas considérés comme étant en compétition.
  - Une bonne sensibilité de cartographie et de définition de variants donne des lectures appariées à un contig ALT, mais qui diffèrent fortement de l'assemblage primaire.

Tri

Le système de cartographie et d'alignement produit un fichier BAM trié par séquence et position de référence par défaut. La création de ce fichier BAM élimine généralement la nécessité d'exécuter la commande *samtools sort* ou toute autre commande post-traitement équivalente. L'option *--enable-sort* peut être utilisée pour activer ou désactiver la création du fichier BAM, comme suit :

- Pour l'activer, mettez la valeur « true » (activée).
- Pour la désactiver, mettez la valeur « false » (désactivée).

Sur le matériel de référence, une analyse effectuée avec la fonction de tri activée augmente le temps d'analyse d'un génome entier 30x d'environ 6 à 7 minutes.

Marquage des répétitions

Le marquage ou la suppression des répétitions de lectures alignées est une pratique exemplaire courante en séquençage de génome entier. Le non-respect de cette pratique peut causer des définitions de variants biaisées et de mauvais résultats.

Le système DRAGEN peut marquer ou supprimer les lectures répétées. Il produit un fichier BAM contenant les répétitions marquées comme telles dans le champ FLAG, ou sans les répétitions.

Durant le test, l'activation du marquage des répétitions augmente seulement un peu le temps de production d'un fichier BAM trié. Le temps supplémentaire n'est que de 1 à 2 minutes pour un génome humain entier 30x, une amélioration importante par rapport aux longs temps d'exécution des outils à source ouverte.

Algorithme de marquage des répétitions

L'algorithme de marquage des répétitions de la plateforme DRAGEN est conçu selon le modèle de la fonction MarkDuplicates de l'outil Picard. Toutes les lectures alignées sont regroupées en sous-ensembles. Les membres de chaque sous-ensemble sont des répétitions potentielles.

Pour que deux paires soient des répétitions, elles doivent présenter :

- Des coordonnées (la position ajustée de la base coupée présente ou absente de l'alignement CIGAR) d'alignement identiques aux deux extrémités.
- Des orientations (directions des deux extrémités, en commençant par la gauche) identiques.

De plus, une lecture non appariée peut être marquée comme une répétition si ses coordonnées et son orientation sont les mêmes que celles de l'extrémité d'une autre lecture, qu'elle soit appariée ou non.

Les alignements non cartographiés ou secondaires ne sont jamais marqués comme des répétitions.

Lorsque la plateforme DRAGEN a identifié un groupe de répétitions, elle détermine quel alignement est le meilleur et marque les autres comme « BAM PCR » ou « optical duplicate » (0x400, ou 1024 en décimales). Pour cette comparaison, le score des répétitions est basé sur la qualité Phred de la séquence moyenne. Les paires reçoivent la somme des scores des deux extrémités, alors que les lectures non appariées reçoivent le score de l'extrémité cartographiée. L'idée derrière ce score est d'essayer, alors que toutes les autres caractéristiques sont équivalentes, de préserver les lectures dont les définitions de bases ont la meilleure qualité.

Si deux lectures (ou paires) ont exactement les mêmes scores de qualité, la plateforme DRAGEN sélectionne la paire ayant le score d'alignement le plus élevé. Si plusieurs paires se retrouvent dans la même situation, la plateforme en choisit une de manière arbitraire.

Le score d'une lecture R non appariée correspond au score de qualité Phred par base, calculé comme suit :

$$score\,(R) = \frac{\sum_i (R.QUAL[i] \text{ where } R.QUAL[i] \geq dedup\_min\_qual)}{sequence\_length(R)}$$

Lorsque R est un enregistrement BAM, la valeur de QUAL correspond aux scores de qualité Phred et l'option de configuration *dedup-min-qual* de la plateforme DRAGEN présente une valeur par défaut de « 15 ». Dans le cas d'une paire, le score est la somme des scores des deux extrémités.

Ce score est stocké comme nombre d'un octet, avec des valeurs arrondies au quart près. Cet arrondissement peut mener à différentes marques de répétitions choisies par Picard, mais puisque les lectures soient très proches les unes des autres au niveau de la qualité, cela a une incidence négligeable sur les résultats de définitions de variants.

Limites du marquage des répétitions

Les limites de la mise en œuvre du marquage des répétitions de la plateforme DRAGEN sont les suivantes :

- Lorsque deux répétitions de lectures ou de paires présentent des scores de qualité de séquences Phred très rapprochés, la plateforme DRAGEN peut choisir un gagnant différent de celui choisi par l'outil Picard. Ces différences ont une incidence négligeable sur les résultats de définitions de variants.
- Le programme exécutable de la plateforme DRAGEN accepte un seul identifiant de librairie comme argument de ligne de commande (PGLB). Pour cette raison, les fichiers d'entrée FASTQ doivent déjà être séparés par identifiant de librairie. Ces identifiants ne peuvent pas être utilisés comme critère de distinction des répétitions et des non-répétitions.

Paramètres de marquage des répétitions

Les options suivantes peuvent être utilisées pour configurer le marquage de répétitions dans la plateforme DRAGEN :

- *--enable-duplicate-marking*

  La valeur « true » (activée) permet d'activer le marquage de répétitions. Lorsque l'option *--enable-duplicate-marking* est activée, le fichier de sortie est trié, peu importe la valeur de l'option *enable-sort*.

- *--remove-duplicates*

  La valeur « true » (activée) supprime les enregistrements de répétitions. Lorsque la valeur est « false » (désactivée), mettez la marque « 0x400 » s'affiche dans le champ FLAG des enregistrements BAM des répétitions. Lorsque l'option *--remove-duplicates* est activée, l'option *enable-duplicate-marking* l'est automatiquement aussi.

- *--dedup-min-qual*

  Spécifie le score de qualité Phred sous lequel une base doit être exclue du calcul du score de qualité utilisé pour faire un choix parmi les répétitions de lectures.

Définition de petits variants

La fonction de définition de petits variants de la plateforme DRAGEN est une fonction de définition d'haplotypes à haute vitesse intégrée à une combinaison matériel-logiciel. La fonction de définition effectue l'assemblage des mutations *de novo* situées dans les régions d'intérêt afin de générer des haplotypes candidats. Elle effectue ensuite des calculs de probabilités sur la lecture à l'aide du modèle de Markov caché (MMC).

La définition de variants est désactivée par défaut. Vous pouvez activer la définition de variant en mettez à l'option *--enable-variant-caller* la valeur « true » (activée).

Algorithme de la fonction de définition de variants

La fonction de définition d'haplotypes de la plateforme DRAGEN effectue les étapes suivantes :

Identification de régions actives – Les zones dans lesquelles plusieurs lectures ne correspondent pas à la référence sont identifiées et les fenêtres qui les entourent (les régions actives) sont sélectionnées aux fins de traitement.

Assemblage d'haplotypes localisé – Dans chaque région active, l'ensemble des lectures qui se chevauchent sont rassemblées dans un graphe de De Bruijn, une représentation graphique basée sur les k-mers (sous-séquences de longueur k) se chevauchant dans chaque lecture ou dans plusieurs lectures. Si toutes les lectures sont identiques, le graphe sera linéaire. Lorsqu'il y a des différences, le graphe forme des bulles avec les chemins qui se rejoignent et s'éloignent. Si la séquence locale est trop répétitive et que la valeur K est trop petite, des cycles peuvent se former, ce qui rend le graphe invalide. Les valeurs K de 10 et 25 sont testées par défaut. Si ces valeurs produisent un graphe non valide, alors des valeurs supplémentaires de 35, 45, 55, 65 sont testées jusqu'à l'obtention d'un graphe sans cycles. À partir de ce graphe de De Bruijn, chaque chemin possible est extrait pour produire une liste exhaustive d'haplotypes candidats (c.-à-d. des hypothèses quant à la véritable séquence d'ADN sur au moins un brin).

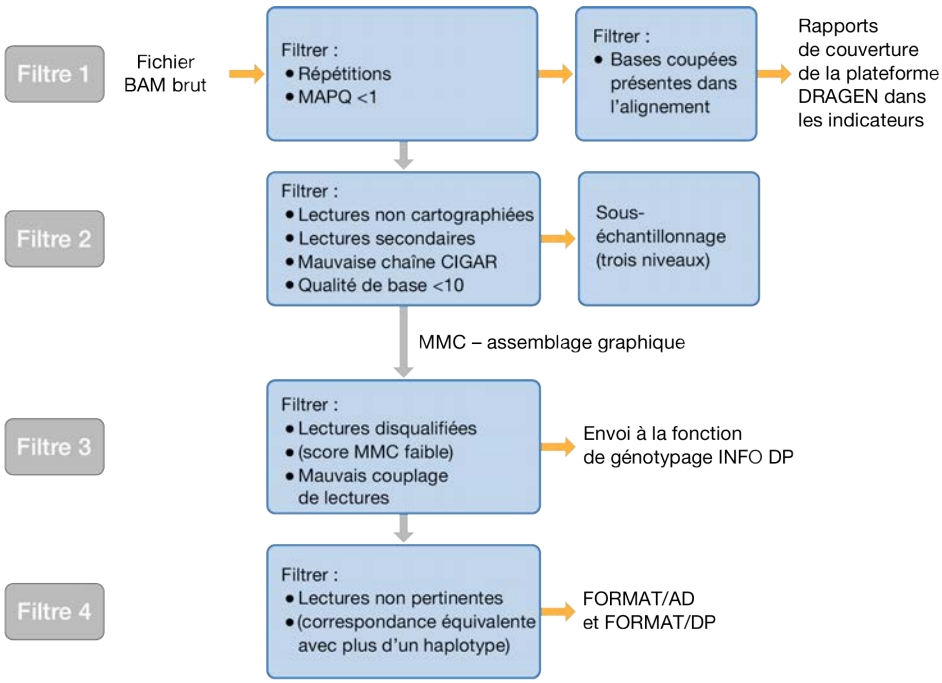
Alignement des haplotypes – Chaque haplotype extrait est aligné avec le génome de référence selon l'algorithme Smith-Waterman, de manière à déterminer quelles variations sont remarquables par rapport à la référence.

Calcul des probabilités des lectures – Chaque lecture est comparée à chaque haplotype pour estimer la probabilité de pouvoir observer la lecture en supposant que l'haplotype est le véritable échantillon d'ADN original. Ce calcul est effectué en évaluant chaque paire selon le modèle de Markov caché (MMC), qui prend en compte les multiples possibilités de modifications de l'haplotype par la PCR ou par des erreurs de séquençage dans la lecture observée. L'évaluation selon le MMC fait appel à une méthode de programmation dynamique pour calculer la probabilité totale d'une série de transitions d'états Markov jusqu'à la lecture observée.

Génotypage – Les combinaisons diploïdes possibles des événements de variants sont formées à partir des haplotypes candidats, et pour chacune d'entre elles, une probabilité conditionnelle d'observation des empilements de lectures est calculée à partir des probabilités d'observation constituantes de chaque lecture selon l'évaluation selon le MMC de chaque haplotype de la paire. La formule bayésienne est ensuite utilisée afin de calculer la probabilité que chaque génotype soit la réalité, en vertu de l'empilement total des lectures observées. Les génotypes présentant la plus grande probabilité sont signalés.

Filtrer les fichiers BAM d'entrée

Les renseignements contenus dans le fichier BAM de l'outil Integrative Genome Viewer (IGV) diffèrent des champs INFO/DP et FORMAT/DP du fichier VCF ou gVCF à cause des filtres appliqués au cours du processus de définition de variants. Quatre filtres sont utilisés pour exclure des lectures des calculs de génotypage. La figure suivante résume ces quatre filtres :



- Le filtre 1 permet de filtrer les lectures suivantes du fichier BAM d'entrée de l'outil IGV :
  - Les répétitions.
  - Les lectures présentant un score MAPQ nul.
  - Les bases coupées présentes dans l'alignement. La plateforme DRAGEN filtre ces bases seulement lors du calcul des rapports de couverture.
  - **[En mode somatique]** Les lectures dont le score MAPQ est inférieur à vc-min-tumor-read-qual lorsque la valeur de vc-min-tumor-read-qual est supérieure à « 1 ».

- Le filtre 2 retranche les bases dont le score de qualité BQ est inférieur à « 10 » et filtre les lectures suivantes :
  - Les lectures non cartographiées.
  - Les lectures secondaires.
  - Les lectures mal alignées.
- Le filtre 3 est appliqué après le sous-échantillonnage et le modèle de Markov caché (MMC). Le filtre 3 permet de filtrer les lectures suivantes :
  - Les lectures mal couplées. Une lecture mal couplée est une lecture cartographiée à deux contigs de référence.
  - Les lectures disqualifiées. Les lectures sont disqualifiées si leur score MMC est inférieur à un certain seuil.
- Le filtre 4 est appliqué après l'exécution de la fonction de génotypage. La fonction de génotypage ajoute des renseignements d'annotation au champ FORMAT. Le filtre 4 permet de filtrer des lectures qui n'apportent aucun renseignement pertinent. Par exemple, si le score MMC d'une lecture est presque égal à celui de deux haplotypes différents, alors la lecture est filtrée, car elle n'apporte pas suffisamment de renseignements pour déterminer lequel des deux haplotypes est le plus probable.
  Le champ INFO/DP comprend à la fois des lectures qui présentent des renseignements pertinents et des lectures qui n'en présentent pas.
  Les champs FORMAT/AD et FORMAT/DP ne comprennent que des lectures qui présentent des renseignements pertinents.

Options de la fonction de définition de variants

Les options suivantes contrôlent l'étape de la fonction de définition de variants du logiciel hôte de la plateforme DRAGEN.

- *--enable-variant-caller*

  Mettez à l'option *--enable-variant-caller* la valeur « *true* » (activée) pour activer l'étape de la fonction de définition de variants du pipeline de la plateforme DRAGEN.

- *--vc-target-bed*

  Il s'agit d'une ligne de commande facultative qui restreint le traitement de la fonction de définition de petits variants, de la couverture du fichier BED de cibles et des indicateurs de définissabilité à des régions spécifiées dans un fichier BED.

  Le fichier texte *--vc-target-bed* contient au moins trois colonnes séparées par des tabulateurs. Les trois premières colonnes sont « chromosome », « start position » (position de départ) et « end position » (position de fin), respectivement. Les positions sont des coordonnées commençant par zéro.

  Par exemple :

  ```
  # header information
    chr11 0 246920
    chr11 255660 255661
  ```

- *--vc-target-bed-padding*

  Cette ligne de commande facultative peut être utilisée pour élargir toutes les régions cibles du fichier BED avec la valeur spécifiée (facultatif). Par exemple, si une région de fichier BED est « 1:1000-2000 » et qu'une valeur d'élargissement de « 100 » est utilisée, cela revient à utiliser une région fichier BED de « 1:900-2100 » et une valeur d'élargissement nulle. Toute valeur d'élargissement ajoutée à l'option *--vc-target-bed-padding* est utilisée par la fonction de définition de petits variants et par les fichiers BED de rapports sur la couverture et la définissabilité de régions cibles.
  La valeur d'élargissement par défaut est « 0 ».

- *--vc-sample-name*

  L'option *--vc-sample-name* est vouée à disparaître. En mode germinal de bout en bout (avec un fichier FASTQ d'entrée), la fonction de définition de petits variants utilise la valeur de RGSM comme nom d'échantillon. En mode somatique sur toute la longueur, l'option *--RGSM-tumor* peut être utilisée pour spécifier le nom de l'échantillon de tumeur. En mode autonome (avec un fichier d'entrée BAM), la fonction de définition de petits variants utilise la valeur de RGSM de l'en-tête du fichier BAM comme nom d'échantillon. En mode somatique, la valeur de RGSM d'un fichier BAM de tumeur est utilisé comme nom d'échantillon de tumeur.

- *--vc-target-coverage*

  L'option *--vc-target-coverage* spécifie la couverture cible aux fins de sous-échantillonnage. La valeur par défaut est « 500 » pour le mode germinal et « 50 » pour le mode somatique.

- *--vc-enable-gatk-acceleration*

  Si la valeur de l'option *--vc-enable-gatk-acceleration* est « true » (activée), la fonction de définition de variants est exécutée en mode GATK (en concordance avec GATK 3.7 en mode germinal et GATK 4.0 en mode somatique).

- *--vc-remove-all-soft-clips*

  Si la valeur de l'option *--vc-remove-all-soft-clips* est « true » (activée), la fonction de définition de variants n'utilise pas de bases coupées présentes dans l'alignement des lectures pour déterminer les variants.

- *--vc-decoy-contigs*

  L'option *--vc-decoy-contigs* spécifie une liste de contigs séparés par des virgules à sauter pendant la définition de variants. Cette option peut être définie dans le fichier de configuration.

- *--vc-enable-decoy-contigs*

  Si la valeur de l'option *--vc-enable-decoy-contigs* est « true » (activée), la définition de variants est activée sur les contigs de leurres. La valeur par défaut est « false » (désactivée).

- *--vc-enable-phasing*

  L'option *–vc-enable-phasing* permet d'activer la mise en phase de variants lorsque cela est possible. La valeur par défaut est « true » (activée).

Options de sous-échantillonnage de la définition de petits variants

Les options de sous-échantillonnage de lectures dans le pipeline de définition de petits variants sont les suivantes :

- *--vc-target-coverage* – Spécifie le nombre maximal de lectures dont la position de départ chevauche une position donnée.
- *--vc-max-reads-per-active-region* – Spécifie le nombre maximal de lectures couvrant une région active donnée.
- *--vc-max-reads-per-raw-region* – Spécifie le nombre maximal de lectures couvrant une région brute donnée.
- *--vc-min-reads-per-start-pos* – Spécifie le nombre minimal de lectures dont la position de départ chevauche une position donnée.

Pour la définition de petits variants mitochondriaux, les options de sous-échantillonnage peuvent être configurées séparément, car les contigs mitochondriaux présentent une plus grande profondeur que le reste des contigs d'un ensemble de données de séquençage de génome entier. Les options suivantes permettent de sous-échantillonner un contig mitochondrial :

- *--vc-target-coverage-mito*
- *--vc-max-reads-per-active-region-mito*
- *--vc-max-reads-per-raw-region-mito*

Les options de couverture cible et de nombres minimaux ou maximaux de lectures dans des régions brutes ou actives ne sont pas directement liées et pourraient être déclenchées de manière indépendante.

L'option de couverture cible est exécutée en premier et vise à limiter le nombre de lectures ayant la même position de départ à une position donnée. Elle ne limite pas la couverture totale à une position donnée.

Le tableau suivant montre les valeurs de sous-échantillonnage par défaut pour chaque mode de définition de petits variants :

Mode	Option de sous-échantillonnage	Valeur par défaut
Germinal	*--vc-target-coverage*	500
Germinal	*--vc-max-reads-per-active-region*	10 000
Germinal	*--vc-max-reads-per-raw-region*	30 000
Somatique	*--vc-target-coverage*	50
Somatique	*--vc-max-reads-per-active-region*	10 000
Somatique	*--vc-max-reads-per-raw-region*	30 000
Couverture élevée	*--vc-target-coverage*	100 000
Couverture élevée	*--vc-max-reads-per-active-region*	200 000
Couverture élevée	*--vc-max-reads-per-raw-region*	200 000
Mitochondrial	*--vc-target-coverage-mito*	40 000
Mitochondrial	*--vc-max-reads-per-active-region-mito*	40 000
Mitochondrial	*--vc-max-reads-per-raw-region-mito*	40 000

L'exemple suivant montre que la profondeur (valeur DP) rapportée dans un enregistrement de variant peut excéder de beaucoup la valeur par défaut de l'option *--vc-target-coverage*, soit « 500 » en mode germinal :

Par exemple, disons que la valeur par défaut de l'option *--vc-target-coverage* est « 500 ». Si 400 lectures commencent à la position 1, que 400 autres lectures commencent à la position 2 et que 400 autres lectures commencent à la position 3, alors l'option de couverture cible n'est pas déclenchée (puisque 400 est inférieur à 500). Si un variant se trouve à la position 4, sa profondeur rapportée peut être aussi élevée que « 1200 ». Cet exemple montre que la profondeur (valeur DP) rapportée dans un enregistrement de variant peut excéder de beaucoup la valeur par défaut de l'option *--vc-target-coverage* :

Après l'étape de couverture cible, le nombre maximal de lectures partageant la même position est « 500 » (si l'option *--vc-target-coverage* est spécifiée à « 500 »).

La prochaine étape du sous-échantillonnage est d'appliquer les limites des options *--vc-max-reads-per-raw-region* et *--vc-max-reads-per-active-region*. Dans cette étape, le nombre maximal de lectures partageant la même position peut être encore plus réduit que la valeur maximale de « 500 » de la première étape. Ces options sont utilisées pour limiter le nombre total de lectures dans une région entière à l'aide d'une méthode de sous-échantillonnage.

Le mécanisme de sous-échantillonnage analyse chaque position à partir de la limite de départ de la région et élimine une lecture à cette position, puis passe à la prochaine position jusqu'à ce que le nombre total de lectures se trouve en deçà du seuil. Plusieurs passages peuvent potentiellement être nécessaires pour que le nombre total de lectures de la région entière se retrouve en deçà du seuil. Une fois que le seuil est respecté, l'étape de sous-échantillonnage est arrêtée, peu importe quelle position a été prise en considération comme étant la dernière dans la région.

Lorsque le nombre de lectures d'une position de départ est inférieur ou égal à la valeur de l'option *--vc-min-reads-per-start-pos*, cette position est sautée de manière à assurer un nombre minimal de lectures (selon l'option *--vc-min-reads-per-start-pos*) à chaque position de départ.

Lorsque le sous-échantillonnage est exécuté, le choix des lectures à garder ou à retirer est aléatoire. Toutefois, le nombre généré aléatoirement devient une valeur par défaut afin d'assurer la production d'un même ensemble de valeurs à chaque analyse. Cela permet d'assurer des résultats exactement reproductibles, ce qui signifie qu'il n'y a aucune variation d'une analyse à l'autre lorsque les mêmes données d'entrée sont utilisées.

Mise en phase et variants mis en phase

La plateforme DRAGEN peut produire des enregistrements de variants mis en phase dans le fichier VCF germinal et le fichier gVCF. Quand deux variants ou plus sont mis en phases, les renseignements relatifs à la mise en phase sont codés dans l'annotation au niveau de l'échantillon, dans le champ FORMAT/PS, qui permet d'identifier dans quel ensemble se trouve le variant mis en phase. La valeur de ce champ est un nombre entier qui représente la position du premier variant mis en phase de l'ensemble. Tous les enregistrements du même contig qui ont une valeur d'ensemble de phase (PS) identique font partie du même ensemble.

```
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="Physical phasing ID
   information, where each unique ID within a given sample (but not
   across samples) connects records within a phasing group">
```

L'exemple suivant montre un fichier gVCF uniéchantillon de la plateforme DRAGEN où deux polymorphismes mononucléotidiques sont mis en phase.

```
chr1 1947645 .       C T,<NON_REF> 48.44 PASS
DP=35;MQ=250.00;MQRankSum=4.983;ReadPosRankSum=3.217;FractionInformativeReads=1.
000;R2_5P_bias=0.000 GT:AD:AF:DP:F1R2:F2R1:GQ:PL:SPL:ICNT:GP:PRI:SB:MB:PS
0|1:20,15,0:0.429:35:9,7,0:11,8,0:47:83,0,50,572,758,622:255,0,255:19,0:4.844e+0
1,8.387e-
05,5.300e+01,4.500e+02,4.500e+02,4.500e+02:0.00,34.77,37.77,34.77,69.54,37.77:11
,9,10,5:12,8,8,7:1947645


chr1 1947648 .       G A,<NON_REF> 50.00 PASS
DP=36;MQ=250.00;MQRankSum=5.078;ReadPosRankSum=2.563;FractionInformativeReads=1.
000;R2_5P_bias=0.000 GT:AD:AF:DP:F1R2:F2R1:GQ:PL:SPL:ICNT:GP:PRI:SB:MB:PS
1|0:16,20,0:0.556:36:8,9,0:8,11,0:48:85,0,49,734,613,698:255,0,255:16,0:5.000e+0
1,7.067e-
05,5.204e+01,4.500e+02,4.500e+02,4.500e+02:0.00,34.77,37.77,34.77,69.54,37.77:10
,6,11,9:8,8,12,8:1947645
```

Pendant l'étape du génotypage, tous les haplotypes et les variants d'une région active sont pris en considération. Si les deux variants d'une paire de variants se trouvent sur les mêmes haplotypes ou si un des variants est homozygote, ils sont mis en phase. Si les variants se trouvent sur des haplotypes différents, ils sont mis en phase opposée. S'il y a des variants hétérozygotes seulement sur certains des mêmes haplotypes, la mise en phase est interrompue et aucun renseignement relatif à la mise en phase ne sera produit pour la région active.

Prise en charge de la ploïdie

La fonction de définition de petits variants ne prend actuellement en charge que la ploïdie 1 ou 2 sur tous les contigs de la référence, à l'exception du contig mitochondrial, qui utilise une approche de fréquence allélique continue (consultez la section *Définition de variants du chromosome mitochondrial* à la page 35). La sélection de la ploïdie 1 ou 2 de tous les autres contigs est déterminée comme suit.

- Si l'option *--sample-sex* n'est pas spécifiée dans la ligne de commande, la fonction d'estimation de la ploïdie détermine le sexe. Si la fonction d'estimation de la ploïdie ne réussit pas à le faire, tous les contigs sont traités avec la ploïdie 2.
- Si l'option *--sample-sex* est spécifiée dans la ligne de commande, les contigs sont traités comme suit.
  - Pour les échantillons de sexe féminin, la plateforme DRAGEN traite tous les contigs avec la ploïdie 2 et ajoute un filtre PloidyConflict aux définitions de variants sur le chromosome Y.
  - Pour les échantillons de sexe masculin, la plateforme DRAGEN traite tous les contigs avec la ploïdie 2, sauf les chromosomes sexuels. La plateforme DRAGEN traite le chromosome X avec la ploïdie 1, sauf dans les régions pseudo-autosomiques où il est entièrement traité avec la ploïdie 2. Le chromosome Y est traité avec la ploïdie 1.

La plateforme DRAGEN détecte les chromosomes grâce à la convention de nommage, soit X/Y ou chrX/chrY. Aucune autre convention de nommage n'est prise en charge.

--sample-sex	Estimation de la ploïdie	Sexe de l'échantillon dans la définition de petits variants
Masculin	Non pertinent	Masculin
Féminin	Non pertinent	Féminin
Non spécifié	XY	Masculin
Non spécifié	XX	Féminin
Non spécifié	Toutes les autres valeurs	Aucun

Définition de variants du chromosome mitochondrial

Aux fins de traitement de la définition de variant du chromosome mitochondrial (chrM), des modifications importantes ont été apportées à partir de la version 3.2 de la plateforme DRAGEN.

Dans les versions précédentes, le chrM était traité soit comme un diploïde (si aucun sexe n'était entré dans la ligne de commande) ou comme un haploïde (si le sexe était entré dans la ligne de commande). En raison de la nature du chromosome M, aucun de ces deux modèles n'était adéquat le fait qu'une cellule donnée possède plusieurs copies du chromosome mitochondrial haploïde et que ces copies ne partagent pas exactement la même séquence d'ADN. Normalement, il y a environ 100 mitochondries dans chaque cellule mammalienne, et chacune de ces mitochondries abrite entre 2 et 10 copies d'ADN mitochondrial (ADNmt). Par exemple, si 20 % des copies chrM ont un variant, alors la fréquence allélique est de 20 %. C'est ce que l'on appelle aussi la fréquence allélique continue. On s'attend à ce que la fréquence allélique des variants chrM se trouve entre 0 % et 100 %.

La plateforme DRAGEN traite maintenant le variant chrM à l'aide d'un pipeline de fréquence allélique continue. Dans ce cas, un seul allèle ALT est pris en considération et la fréquence allélique est estimée, pouvant se trouver n'importe où entre 0 % et 100 %.

Le traitement du chromosome mitochondrial est maintenant plus exact que dans les versions précédentes puisque, dans les versions antérieures à 3.2 de la plateforme DRAGEN, la définition de faibles fréquences alléliques ne produisait pas de résultats (puisqu'on s'attendait à ce que la fréquence se rapproche des 100 % dans un modèle haploïde). Dans la version actuelle, il est possible de voir chacun des variants de l'allèle ALT du chromosome mitochondrial, sur l'ensemble des fréquences alléliques (de faible à élevée).

La qualité du variant (QUAL) n'est pas ajoutée dans les enregistrements de variant chrM. Le score de confiance est plutôt INFO/LOD,

```
##INFO=<ID=LOD,Number=1,Type=Float,Description="Variant LOD score">
```

qui exprime la confiance qu'un variant soit présent à un locus donné.

La qualité du génotype (GQ) n'est pas ajoutée dans les enregistrements de variants chrM, puisque la plateforme DRAGEN n'effectue pas de tests visant plusieurs génotypes diploïdes candidats. À la place, un allèle ALT est considéré comme étant un variant candidat. Si la valeur du champ INFO/LOD est supérieur à la valeur de l'option vc-lod-call-threshold (la valeur par défaut est « 4 »), alors « 0/1 » est figé dans le code du champ FORMAT/GT, et le champ FORMAT/AF offre une estimation de la fréquence allélique du variant dans l'étendue [0,1].

Les filtres suivants peuvent être appliqués à la définition de variants chrM.

--vc-lod-call-threshold

La limite de détection pour l'ajout de définitions dans le fichier VCF. La valeur par défaut est « 4 ».

--vc-lod-filter-threshold

La limite de détection pour marquer les définitions de fichiers VCF ajoutés comme filtrées. La valeur par défaut est « 6,3 ».

Si la valeur du champ INFO/LOD est inférieure à la valeur de l'option vc-lod-call-threshold, le variant n'est pas ajouté dans le fichier VCF.

Si la valeur du champ INFO/LOD est supérieure à la valeur de l'option vc-lod-call-threshold, mais inférieure à la valeur de l'option vc-lod-filter-threshold, le variant est ajouté dans le fichier VCF, mais la colonne FILTER (filtre) indique lod_fstar.

Si la valeur du champ INFO/LOD est supérieure à la valeur de l'option vc-lod-call-threshold et de l'option vc-lod-filter-threshold, le variant est ajouté dans le fichier VCF et la colonne FILTER (filtre) indique « PASS » (réussite).

Les exemples suivants montrent des enregistrements VCF de variants chrM. L'une des définitions a une fréquence allélique très élevée, l'autre très faible. Dans les deux cas, la valeur du champ FORMAT/LOD est supérieure à la valeur de l'option emit_threshold. Si le champ FORMAT/LOD affiche aussi un seuil supérieur à celui du filtre lod_fstar, alors l'annotation du filtre est « PASS » (réussite).

```
chrM 2259 . C T . PASS
   DP=9791;MQ=60.00;LOD=38838.40;FractionInformativeReads=0.994
   GT:AD:AF:F1R2:F2R1:DP:SB:MB
   0/1:5,9729:0.999:1,5007:4,4722:9734:3,2,4885,4844:1,4,4807,4922
chrM 16192 . C T . PASS
   DP=9644;MQ=60.00;LOD=26.12;FractionInformativeReads=0.992
   GT:AD:AF:F1R2:F2R1:DP:SB:MB
   0/1:9537,26:0.003:5530,16:4007,10:9563:4484,5053,13,13:4961,4576,19,
```

Modes VFC combiné et gVCF

En mode gVCF, qui est disponible pour le pipeline germinal, les régions NON_REF sont affichées sur les enregistrements de variants. L'exemple suivant montre des régions NON_REF et des régions de variants pour le fichier de sortie du chrM en mode gVCF :

```
chrM 751 . A <NON_REF> . PASS END=1437 GT:AD:DP:GQ:MIN_DP:PL:SPL:ICNT
   0/0:6920,9:6929:99:4077:0,120,1800:0,255,255:40,4
chrM 1438 . A G,<NON_REF> . PASS
   DP=8500;MQ=57.39;LOD=30441.87;FractionInformativeReads=0.871
   GT:AD:AF:F1R2:F2R1:DP:SB:MB
   0/1:0,7400,0:1.000:0,3765,0:0,3635,0:7400:0,0,3994,3406:0,0,3633,3767
chrM 1439 . A <NON_REF> . PASS END=2258 GT:AD:DP:GQ:MIN_DP:PL:SPL:ICNT
   0/0:6120,10:6130:99:4190:0,120,1800:0,255,255:40,14
```

FORMAT/GT

Dans les régions NON_REF, la valeur de FORMAT/GT est fixée à « 0/0 » et au niveau du locus du variant, à « 0/1 ».

La valeur de FORMAT/GT pour un chrM n'est pas déterminée par la valeur de FORMAT/AF, mais plutôt par si le variant est produit à une certaine position ou non. La décision de produire le variant ou non est déterminée par la comparaison du score INFO/LOD et d'un seuil. Tous les variants pour lesquels la valeur de FORMAT/LOD est inférieure au seuil (emit_threshold) sont produits. Les variants dont la valeur de FORMAT/LOD est inférieure au seuil (emit_threshold) ne sont pas produits dans le fichier gVCF et sont placés en alternance dans une région NON_REF avec une valeur FORMAT/GT de « 0/0 ».

L'une des deux situations suivantes est possible à n'importe quelle position donnée.

- Un variant est détecté par la fonction de génotypage à une position donnée et la valeur de FORMAT/LOD est supérieure au seuil (emit_threshold). Dans ce cas, la valeur de FORMAT/GT est de « 0/1 » et la plateforme DRAGEN produit des valeurs FORMAT/AD, FORMAT/DP et FORMAT/AF pour cette position.
- Aucun variant n'est détecté par la fonction de génotypage à une position donnée ou un variant est détecté, mais la valeur de FORMAT/LOD est inférieure au seuil (emit_threshold). Dans ce cas, la valeur de FORMAT/GT est de « 0/0 » et la position est placée en alternance avec des positions contiguës si elles se trouvent dans la même situation. Toutes les positions présentant une valeur de FORMAT/GT de « 0/0 » sont placées en alternance les unes avec les autres. La valeur FORMAT/DP rapportée pour la bande est calculée comme étant la profondeur médiane de toutes les positions de la bande. Les valeurs FORMAT/AD rapportées dans la bande sont les valeurs sélectionnées à la position où la valeur de FORMAT/DP est égale à la profondeur médiane. Dans le fichier VCF combiné, la valeur de FORMAT/AF est calculée à partir de la valeur de FORMAT/AD.

Les exemples suivants montrent un enregistrement de variant chrM dans un trio VCF combiné.

Le premier échantillon montre un variant, mais la valeur de FORMAT/FT est « lod_fstar » parce que la valeur de FORMAT/LOD est inférieure au seuil lod_fstar_threshold.

Le deuxième échantillon ne contient pas de variant et la valeur de FORMAT/AF est nulle.

Le troisième échantillon ne contient pas de variant même si la valeur de FORMAT/AF est supérieure à « 0 », ce qui signifie que cette position, pour l'échantillon, est une région NON_REF dans laquelle aucun variant n'a été détecté avec un score de confiance suffisamment élevé.

```
chrM 2622 . G A . . DP=11199;MQ=59.73 GT:AD:AF:DP:FT:LOD:F1R2:F2R1
   0/1:3375,8:0.002:3383:lod_fstar:4.4:1689,2:1686,6
   0/0:5629,1:0:5118:PASS:.:.:. 0/0:3505,8:0.003:2645:PASS:.:.:.
```

Formules de QUAL, QD et GQ

Dans les fichiers VCF et gVCF uniéchantillon, QUAL suit la définition de la spécification VCF (https://samtools.github.io/hts-specs/VCFv4.3.pdf).

- QUAL est la probabilité de l'échelle Phred que le site n'ait aucun variant. Sa valeur est calculée comme suit :

  ```
  QUAL = -10*log10 (posterior genotype probability of a homozygous-
     reference genotype (GT=0/0))
  ```

  C'est-à-dire, QUAL = GP (GT = 0/0), où GP est la probabilité de génotype a posteriori dans l'échelle Phred.

  Une valeur de QUAL de 20 signifie qu'il y a une probabilité de 99 % qu'il y ait un variant au site. Les valeurs de GP sont également données à l'échelle Phred dans le fichier VCF.

- GQ est la probabilité de l'échelle Phred que la définition soit incorrecte.

  GQ = -10 * log 10(p), où p est la probabilité que la définition soit incorrecte.

  GQ = -10 * log 10(somme(10.^(-GP(i)/10))) où la somme est choisie plutôt que génotype qui n'a pas gagné.

  Donc, une valeur de GQ de 3 signifie qu'il y a une probabilité de 50 pour cent que la définition soit incorrecte et une valeur de GQ de 20 signifie qu'il y a une probabilité de 1 pour cent que la définition soit incorrecte.

- QD est la valeur de QUAL normalisée en fonction de la profondeur de lecture, DP.

Indicateur	QUAL	GQ	QD
Description	Probabilité que le site n'ait aucun variant	Probabilité que la définition soit incorrecte	QUAL normalisée en fonction de la profondeur
Formule	QUAL = GP(GT=0/0)	GQ=-10*log10(p)	QUAL/DP
Échelle	Phred non signée	Phred non signée	Phred non signée
Exemple avec des nombres	QUAL = 20 : probabilité de 1 % qu'il n'y ait pas de variant à ce site QUAL = 50 : probabilité de 1 sur 1e5 qu'il n'y ait pas de variant à ce site	GQ = 3, probabilité de 50 % que la définition soit incorrecte GQ = 20, probabilité de 1 % que la définition soit incorrecte	

Modification de la plage de valeurs entre les versions précédentes de la plateforme DRAGEN/GATK et la plateforme DRAGEN 2.6+

La plage des valeurs de qualité du variant (QUAL, et par conséquent la qualité du génotype GQ et la qualité basée sur la profondeur QD) a été modifiée entre les versions antérieures de la plateforme DRAGEN (et/ou GATK) et la version 2.6+. Dans les versions antérieures de la plateforme DRAGEN, les calculs de la qualité du variant suivaient l'approche de GATK. Dans les versions récentes de la plateforme DRAGEN (à partir de la version 2.6), l'algorithme de détection de petits variants a été amélioré afin d'offrir des valeurs de qualité du variant (QUAL, et par conséquent GQ et QD) plus réalistes (c.-à-d. des valeurs plus petites que dans GATK).

Les valeurs QUAL du variant ont été modifiées parce que les calculs de probabilités utilisés par GATK et les versions antérieures de la plateforme DRAGEN supposaient qu'il n'y avait aucune corrélation entre les erreurs des différentes lectures. Cela comprend les erreurs de cartographie et les erreurs de définition de bases. Dans un monde réel, où les erreurs sont corrélées, cela mène à un gonflement des scores QUAL ainsi qu'à des performances de détection non optimales. Les algorithmes de la plateforme DRAGEN 2.6+ intègrent l'hypothèse de la corrélation entre les erreurs au sein de la fonction de définition de variants, ce qui mène à une amélioration de l'exactitude de détection et à des scores QUAL plus réalistes. Les scores QUAL plus réalistes dans la plateforme DRAGEN 2.6+ sont inférieurs aux scores QUAL gonflés dans GATK.

Histogramme des scores QUAL, QD et GQ

Le tracé de l'histogramme des valeurs QUAL (et QD et GQ) à partir des définitions d'un fichier VCF de la plateforme DRAGEN 2.6+ peut présenter un pic correspondant à une certaine valeur (p. ex., 50). Une petite portion des valeurs QUAL se trouve en deçà ou au-dessus de ce niveau. Cela est causé par l'algorithme de détection des lectures étrangères (FRD) qui définit une limite pour la valeur du score de qualité QUAL (et indirectement des scores GQ et QD) des variants hétérozygotes. Consultez la section *Détection des lectures étrangères* à la page 40. La valeur exacte dépend de la cartographie entre le score MAPQ maximal de la fonction de cartographie et un score de confiance Phred indiquant que la lecture est bien cartographiée. Les variants homozygotes ne sont pas assujettis à cette limite, car l'hypothèse de la fonction FRD ne s'applique qu'aux variants hétérozygotes (parce que l'option *--vc-frd-beta-max*, qui spécifie la fréquence allélique maximale pour l'hypothèse relative à la lecture étrangère a une valeur de « 0,5 »).

Cela signifie que le score QUAL des variants hétérozygotes est limité, alors que le score QUAL des variants homozygotes ne l'est pas. Toutefois, l'échelle QUAL est la même pour tous les variants (qu'ils soient hétérozygotes ou homozygotes); seule la limite est différente.

Modélisation des erreurs corrélées dans les lectures

La fonction de définition de variants 2.6 de la plateforme DRAGEN et des versions plus récentes dispose de deux algorithmes pour modéliser les erreurs corrélées dans les lectures d'un empilement donné.

- L'algorithme de détection de lectures étrangères (FRD) permet de détecter les lectures incorrectement cartographiées. L'algorithme FRD modifie le calcul de la probabilité afin de prendre en compte la possibilité qu'un sous-ensemble des lectures soit incorrectement cartographié. Plutôt que d'assumer que les erreurs de cartographie se produisent indépendamment pour chaque lecture, l'algorithme estime la probabilité qu'une suite de lectures a été incorrectement cartographiée en incorporant des données probantes comme des scores MAPQ et des fréquences alléliques biaisées.
- L'algorithme de diminution de la qualité des bases (BQD) permet de détecter les erreurs corrélées de définition de bases : la plateforme DRAGEN dispose d'un mécanisme qui détecte les erreurs de définition de bases systématiques et corrélées qui sont causées par l'instrument de séquençage. Le mécanisme de détection tire parti de certaines propriétés de ces erreurs (biais lié au brin, position de l'erreur dans la lecture, qualité des bases) afin d'estimer la probabilité que les allèles résultent d'événements d'erreur systématique plutôt que d'un véritable variant.

La modélisation des erreurs corrélées fournit des valeurs de score de confiance (QUAL, GQ, QD) qui sont réalistes. Ces valeurs sont beaucoup plus basses que les scores de confiance produites par GATK.

Détection des lectures étrangères

Les fonctions classiques de définition de variants traitent les erreurs de cartographie comme des événements d'erreur indépendants par lecture, ignorant le fait que de telles erreurs surviennent habituellement en rafale. Cela peut produire un score de confiance élevé malgré une cartographie de faible qualité ou une fréquence allélique biaisée. Pour limiter ce problème, les fonctions de définition de variants classiques peuvent comprendre un score MAPQ minimal à inclure dans le calcul, mais ce seuil peut causer le rejet de preuves valides et élimine mal les faux positifs.

La plateforme DRAGEN v2.6 et ses versions subséquentes ont mis en place la fonction de détection des lectures étrangères, une extension de l'algorithme de génotypage existant, en faisant l'hypothèse supplémentaire que certaines lectures de l'empilement sont de nature étrangère (c.-à-d. que leur véritable emplacement est ailleurs dans le génome de référence). L'algorithme exploite de multiples propriétés (fréquence allélique biaisée et cartographie de faible qualité) et ajoute ces données au calcul de la probabilité selon une approche mathématique rigoureuse.

La récupération des faux négatifs, la correction des génotypes et la diminution du seuil de qualité MAPQ pour les saisies de lectures dans la fonction de définition de variants permettent d'accroître la sensibilité. Le retrait des faux positifs et la correction des génotypes permettent d'accroître la spécificité.

La détection des lectures étrangères est une fonction plus puissante que les approches mettant l'accent sur la filtration après VCF pour améliorer la mesure F, puisque, au lieu de simplement détecter des résultats douteux (p. ex., fondés sur la profondeur allélique ou les erreurs de lecture) après la définition des variants, l'algorithme de détection intègre directement la présence de lectures étrangères par une détection rigoureuse en fonction du maximum de vraisemblance.

Diminution de la qualité des bases

Les fonctions classiques de définition de variants sont conçues selon l'hypothèse que les erreurs de séquençage sont indépendantes d'une lecture à l'autre. D'après cette hypothèse, il est peu probable que plusieurs erreurs identiques se produisent dans un locus en particulier.

Toutefois, après analyse des ensembles de données de séquençage nouvelle génération, il a été observé que des rafales d'erreurs sont beaucoup plus fréquentes que ce que l'hypothèse d'indépendance prédit, et ces rafales d'erreurs peuvent provoquer de nombreux faux positifs.

Heureusement, ces erreurs ont des caractéristiques propres, ce qui permet de les différencier des vrais variants. L'algorithme de la diminution de la qualité des bases (BDQ) est un mécanisme de détection qui exploite certaines propriétés de ces erreurs (biais lié au brin, position de l'erreur dans la lecture et faible qualité moyenne des bases dudit sous-ensemble de lectures, au locus d'intérêt) et les intègre au calcul de la probabilité dans la fonction de génotypage.

Fonction de définition de régions d'homozygotie

Les régions d'homozygotie (ROH) sont détectées lors de l'exécution de la fonction de définition de petits variants. La fonction de définition détecte et produit les régions d'homozygotie des définitions du génome entier sur les chromosomes humains autosomiques. Les chromosomes sexuels sont ignorés. La sortie de régions d'homozygotie permet aux outils en aval de dépister et de prédire la consanguinité entre les parents de l'individu.

L'algorithme ROH est exécuté sur les définitions de petits variants. Il exclut les variants avec des sites multialléliques, les indels, les variants complexes, les définitions filtrées parce qu'elles n'avaient pas la valeur « PASS » (réussite) et les sites de référence homozygotes. Les définitions de variants sont alors filtrées à l'aide d'un fichier BED de régions à ignorer, puis filtrées selon la profondeur. La valeur par défaut de la proportion

de définitions filtrées est 0,2, ce qui filtre les 10 % des définitions qui ont les valeurs de profondeur les plus élevées et les 10 % qui ont les valeurs de profondeur les plus basses. Les définitions en résultant sont alors utilisées pour trouver les régions.

Une région est définie comme des définitions de variants consécutives sur le chromosome entre lesquelles il n'y a pas de grand écart. Autrement dit, les régions sont divisées par chromosome ou par grands écarts sans définitions de variant de polymorphisme mononucléotidique. La taille de l'écart est réglée à 3 Mb.

Options relatives aux régions d'homozygotie

- *--vc-enable-roh*

  Active ou désactive la fonction de définition de régions d'homozygotie (ROH) en mettant la valeur « true » (activée) ou « false » (désactivée) à cette option. Option activée par défaut pour les autosomes humains seulement.

- *--vc-roh-blacklist-bed*

  La fonction de définition de ROH ignore les variants qui se trouvent dans toute région d'un fichier BED de régions à ignorer fourni. La plateforme DRAGEN fournit les fichiers de régions à ignorer pour tous les génomes humains les plus utilisés. Le logiciel sélectionne automatiquement une liste de régions à ignorer qui correspond au génome utilisé, sauf si cette option spécifie explicitement un fichier.

Fichier de sortie de ROH

La fonction de définition de ROH produit un fichier de sortie de ROH nommé <output-file-prefix>.roh.bed dans lequel chaque ligne représente une ROH. Le fichier BED contient les colonnes suivantes :

```
Chromosome Start End Score #Homozygous #Heterozygous
```

Où

- Score est une fonction du nombre de variants homozygotes et hétérozygotes où chaque variant homozygote fait augmenter le score d'une valeur prédéfinie et où chaque variant hétérozygote fait baisser le score de (1 - valeur prédéfinie), où la valeur prédéfinie se trouve dans l'intervalle (0, 1).
- Les positions de Start (Début) et de End (Fin) sont des coordonnées commençant par 0, dans un intervalle semi-ouvert.
- #Homozygous est le nombre de variants homozygotes dans la région.
- #Heterozygous est le nombre de variants hétérozygotes dans la région.

La fonction de définition produit également un fichier d'indicateurs nommé <output-file-prefix>.roh_metrics.csv qui contient le nombre de grandes ROH et le pourcentage de polymorphismes mononucléotidiques qui se trouvent dans les grandes ROH (plus de 3 Mb).

Fichier de sortie de fréquence d'allèle B

Le fichier de sortie de fréquence d'allèle B (BAF) est activé par défaut dans les analyses germinales et somatiques de fichier VCF et gVCF.

La valeur de BAF est égale à AF ou à (1 – AF), où

- AF = (alt_count/(ref_count + alt_count))
- BAF = 1 – AF, seulement lorsque la base de référence est inférieure à la base alternative et que l'ordre de priorité pour les bases est A < T <G < C < N

Pour chaque entrée de fichier VCF de petits variants avec exactement un allèle SNP alternatif, la sortie contient une entrée correspondante dans le fichier BAF de sortie.

- Les lignes <NON_REF> sont exclues.
  - Les variants ForceGT (marqués par l'étiquette « FGT » dans le champ INFO) ne sont pas compris dans le fichier de sortie, à moins que le variant ne contienne également l'étiquette « NML » dans le champ INFO.
  - Les variants où les valeurs de ref_count et de alt_count sont toutes les deux zéro ne sont pas compris dans le fichier de sortie.

Options relatives au fichier BAF

--vc-enable-baf

Permet d'activer ou de désactiver le fichier de sortie de fréquence d'allèle B. Cette option est activée par défaut.

Fichier BAF de sortie

Les fichiers générés sont des fichiers compressés bigWig, nommés <output-file-prefix>.baf.bw et <output-file-prefix>.hard-filtered.baf.bw. Les fichiers filtrés contiennent seulement des entrées pour les variants qui correspondent aux filtres définis dans le fichier VCF (c.-à-d. les entrées marquées « PASS » [réussite]).

Chaque entrée comprend les renseignements suivants :

```
Chromosome Start End BAF
```

Où :

- La valeur de Chromosome est une chaîne qui correspond à un contig de référence.
- Les valeurs de Start (Départ) et de End (Fin) sont des intervalles à moitié ouverts et basés sur zéro.
- La valeur de BAF est une valeur de point flottant.

Mode somatique

Le pipeline somatique de la plateforme DRAGEN permet une analyse ultrarapide de données de séquençage nouvelle génération (SNG) afin d'identifier des mutations associées au cancer dans des chromosomes somatiques. La plateforme DRAGEN définit les variants du nombre de copies et les indels à la fois à partir de paires d'échantillons de tumeur et de référence et d'échantillons de tumeur seulement.

Dans le pipeline d'échantillons de tumeur et de référence, les deux échantillons sont analysés de manière combinée afin que les variants germinaux soient exclus, ce qui génère un fichier de sortie propre aux mutations tumorales. Le pipeline d'échantillons de tumeur seulement produit un fichier VCF qui peut être analysé de nouveau afin d'identifier des mutations tumorales. Lorsqu'il s'agit d'un échantillon de tumeur, les deux pipelines n'effectuent aucune hypothèse relative à la ploïdie et permettent la détection d'allèles à fréquence faible.

Les résultats, après avoir traversé une série de filtres, sont produits dans un fichier VCF. Les variants qui ne correspondent pas aux filtres sont conservés dans le fichier VCF de sortie, mais porte une annotation à cet effet dans la colonne FILTER.

Options du mode somatique

Le mode somatique a les options de lignes de commande suivantes :

- *--tumor-fastq1* et *--tumor-fastq2*

  Les options *--tumor-fastq1* et *--tumor-fastq2* sont utilisées pour mettre une paire de fichiers FASTQ en entrée dans les fonctions de cartographie, d'alignement et de définition de variants somatiques. Ces options peuvent être utilisées avec d'autres options FASTQ en mode d'échantillons de tumeur et de référence.

  Par exemple :

  ```
  dragen -f -r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
     --tumor-fastq1 <TUMOR_FASTQ1> \
     --tumor-fastq2 <TUMOR_FASTQ2> \
     --RGID-tumor <RG0-tumor> --RGSM-tumor <SM0-tumor> \
     -1 <NORMAL_FASTQ1> \
     -2 <NORMAL_FASTQ2> \
     --RGID <RG0> -RGSM <SM0> \
     --enable-variant-caller true \
     --output-directory /staging/examples/\
     --output-file-prefix SRA056922_30x_e10_50M
  ```

- *--tumor-fastq-list*

  L'option *--tumor-fastq-list* est utilisée pour mettre une liste de fichiers FASTQ en entrée dans les fonctions de cartographie, d'alignement et de définition de variants somatiques. Cette option peut être utilisée avec d'autres options FASTQ en mode d'échantillons de tumeur et de référence. Par exemple :

  ```
  dragen -f \
     -r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
     --tumor-fastq-list <TUMOR_FASTQ_LIST> \
     --fastq-list <NORMAL_FASTQ_LIST> \
     --enable-variant-caller true \
     --output-directory /staging/examples/\
     --output-file-prefix SRA056922_30x_e10_50M
  ```

- *--tumor-bam-input* et *--tumor-cram-input*

  Les options *--tumor-bam-input* et *--tumor-cram-input* sont utilisés pour mettre un fichier BAM ou CRAM cartographié en entrée dans la fonction de définition de variants somatiques. Ces options peuvent être utilisées avec d'autres options BAM/CRAM en mode d'échantillons de tumeur et de référence.

- *--vc-min-tumor-read-qual*

  L'option *--vc-min-tumor-read-qual* spécifie la qualité minimale (score MAPQ) de lecture à considérer pour la définition de variants. La valeur par défaut de cette option est « 3 » pour une analyse d'échantillons de tumeur et de référence et « 20 » pour une analyse d'échantillon de tumeur seulement.

- *--vc-callability-tumor-thresh* ou *--vc-callability-normal-thresh*

  L'option *--vc-callability-tumor-thresh* spécifie le seuil de définissabilité pour les échantillons de tumeur seulementé. L'option *--vc-callability-normal-thresh* spécifie le seuil de définissabilité pour les échantillons de tumeur et de référence. Le fichier BED de rapport sur les régions somatiques définissables comprend toutes les régions qui sont supérieures au seuil de la tumeur. Pour en savoir plus sur le rapport BED des régions de définition somatique, consultez la section *Rapport sur les régions somatiques définissables* à la page 117.

La valeur par défaut pour le seuil d'échantillon de tumeur seulement est « 15 » et la valeur par défaut pour le seuil des échantillons de tumeur et de référence est « 5 ».

- *--vc-enable-liquid-tumor-mode* et *vc-tin-contam-tolerance*

  Dans une analyse d'échantillons de tumeur et de référence, la plateforme DRAGEN répond à une contamination d'échantillon de tumeur dans l'échantillon de référence (TiN, pour tumor-in-normal) en exécutant le mode de tumeur liquide. Le mode de tumeur liquide est désactivé par défaut. L'option *--vc-enable-liquid-tumor-mode* active le mode de tumeur liquide avec une tolérance TiN de contamination maximale par défaut de « 0,15 ». Si vous utilisez la tolérance TiN de contamination maximale par défaut, il est attendu que des variants somatiques soient observés dans l'échantillon de référence avec des fréquences alléliques, jusqu'à 15 % des allèles correspondants dans l'échantillon de tumeur. L'option *vc-tin-contam-tolerance* active le mode de tumeur liquide et définit une tolérance TiN différente de zéro.

Filtres de définition postsomatiques

Les options suivantes sont disponibles comme filtres de définition postsomatiques :

- *--vc-sq-call-threshold*

  Produit des définitions dans le fichier VCF. La valeur par défaut est « 3 ». Si la valeur de l'option *vc-sq-filter-threshold* est inférieure à celle de l'option *vc-sq-call-threshold*, la valeur de seuil du filtre est utilisée plutôt que la valeur de seuil de la définition.

- *--vc-sq-filter-threshold*

  Marque les définitions VCF produites comme filtrées. La valeur par défaut est « 17,5 » pour le mode relatif aux échantillons de tumeur et de référence et « 6,5 » pour le mode relatif aux échantillons de tumeur seulement.

- *--vc-enable-triallelic-filter*

  Active le filtre multiallélique. La valeur par défaut est « true » (activée).

- *--vc-enable-af-filter*

  Active le filtre de fréquence allélique. La valeur par défaut est « false » (désactivée). Lorsque la valeur est « true » (activée), le fichier VCF exclut les variants dont la fréquence allélique se trouve en deçà du seuil de définition de la fréquence allélique ou les variants ayant une fréquence allélique qui se trouve en deçà du seuil du filtre de la fréquence allélique et avec une étiquette de filtre de fréquence allélique faible. Le seuil de définition de la fréquence allélique est « 1 % » par défaut et le seuil du filtre de la fréquence allélique est « 5 % » par défaut. Pour modifier les valeurs du seuil, utilisez les options de ligne de commande suivantes :

  `vc-af-call-threshold`

  `vc-af-filter-threshold`

- *--vc-enable-non-homref-normal-filter*

  Active le filtre de référence non-homref. La valeur par défaut est « true » (activée). Lorsque la valeur est « true » (activée), le fichier VCF filtre les variants si le génotype de l'échantillon de référence n'est pas homozygote.

Mode somatique	Identifiant du filtre	Description
Échantillons de tumeur seulement et échantillons de tumeur et de référence	clustered_events	Des événements de génération d'amplifiats ont été observés dans une région active donnée. Le seuil pour les événements de génération d'amplifiats est configurable (la valeur par défaut est inférieure ou égale à « 3 »). Filtre activé seulement avec l'option *--vc-enable-gatk-acceleration=true*.
Échantillons de tumeur seulement et échantillons de tumeur et de référence	weak_evidence	Le variant ne respecte pas le seuil de probabilité. Le rapport de probabilité pour le score SQ de l'échantillon de tumeur et de référence est inférieur à « 17,5 » ou inférieur à « 6,5 » pour le score SQ de l'échantillon de tumeur seulement.
Échantillons de tumeur seulement et échantillons de tumeur et de référence	multiallelic	Le site est filtré s'il y a au moins deux allèles ALT à cet emplacement de l'échantillon de tumeur.
Échantillons de tumeur seulement et échantillons de tumeur et de référence	str_contraction	Une erreur PCR est soupçonnée là où l'allèle ALT a une unité répétée de moins que la référence. Filtre activé seulement avec l'option *--vc-enable-gatk-acceleration=true*.
Échantillons de tumeur seulement et échantillons de tumeur et de référence	base_quality	La qualité de base médiane des lectures ALT pour ce locus est inférieure à « 20 ».
Échantillons de tumeur seulement et échantillons de tumeur et de référence	mapping_quality	La qualité de cartographie médiane des lectures ALT pour ce locus est inférieure à « 20 » (échantillons de tumeur et de référence) ou inférieure à « 30 » (échantillon de tumeur seulement).
Échantillons de tumeur seulement et échantillons de tumeur et de référence	fragment_length	La différence absolue entre la longueur de fragment médiane des lectures ALT et des lectures REF à un locus donné est supérieure à « 10 000 ».
Échantillons de tumeur seulement et échantillons de tumeur et de référence	read_position	La médiane des distances entre le début et la fin de la lecture et un locus donné est inférieure à « 5 » (le variant est trop près du bord de toutes les lectures).
Échantillons de tumeur seulement et échantillons de tumeur et de référence	panel-of-normals	Vu dans au moins un échantillon du fichier VCF du panel de référence.
Échantillons de tumeur seulement et échantillons de tumeur et de référence	low_af	La fréquence allélique est en deçà du seuil spécifié grâce à l'option *--vc-af-filter-threshold* (la valeur par défaut est « 5 % »). Filtre activé seulement avec l'option *--vc-enable-af-filter=true*.

Filtres postsomatiques de contrôle de l'échantillon de référence correspondant

Mode somatique	Identifiant du filtre	Description
Échantillons de tumeur et de référence	noisy_normal	Plus de trois allèles sont observés dans l'échantillon de référence à une fréquence allélique supérieure à 9,9 %.
Échantillons de tumeur et de référence	alt_allele_in_normal	La fréquence allélique ALT de l'échantillon de référence est supérieure à « 0,2 » plus la tolérance de contamination maximale. En mode de tumeur solide, la valeur est « 0 ». En mode de tumeur liquide, la valeur par défaut est « 0,15 ».
Échantillons de tumeur et de référence	filtered_reads	Plus de 90 % des lectures ont été filtrées.
Échantillons de tumeur et de référence	no_reliable_ supporting_read	Aucune lecture fiable n'a été trouvée dans l'échantillon de tumeur. Une lecture fiable est une lecture en soutien à l'allèle ALT ayant une qualité de cartographie ≥ 40, une longueur de fragment ≤ 10 000, une qualité de définition de base ≥ 25 et une distance départ/fin de la lecture ≥ 5.
Échantillons de tumeur et de référence	strand_artifact	Important biais lié au brin. Allèle ALT soutenu par au moins quatre lectures, mais seulement un seul brin.
Échantillons de tumeur et de référence	too_few_supporting_ reads	Variant soutenu par < 3 lectures dans l'échantillon de tumeur.
Échantillons de tumeur et de référence	non-homref-normal	Le génotype de l'échantillon de référence n'est pas une référence homozygote.
Échantillons de tumeur et de référence	germline_risk	Probabilité que l'allèle soit présent dans la référence > 0,025. Filtre activé seulement avec l'option *--vc-enable-gatk-acceleration=true*.
Échantillons de tumeur et de référence	artifact_in_normal	La limite de détection de tumeur de l'ensemble de référence (limite de détection d'artéfacts de la référence) est > 0,0. Les artéfacts ne sont pas appelés de référence si la fraction allélique de la référence est significativement plus petite que la fraction allélique de la tumeur (normalAlleleFraction < (0,1 * tumorAlleleFraction). Filtre activé seulement avec l'option *--vc-enable-gatk-acceleration=true*.

La qualité du variant (QUAL) n'est pas produit dans les enregistrements de variant somatique. Le score de confiance est plutôt FORMAT/SQ.

```
##FORMAT=<ID=SQ,Number=1,Type=Float,Description="Somatic quality">
```

Ce champ est propre à l'échantillon. Pour les échantillons de tumeur, il qualifie les données probantes indiquant qu'un variant somatique est présent à un locus donné.

Si un échantillon de référence est aussi disponible, la valeur du champ FORMAT/SQ correspondante quantifie les données probantes indiquant que l'échantillon de référence est une référence homozygote dans un locus donné.

La qualité du génotype (GQ) n'est pas indiquée dans les enregistrements de variant somatique, car la plateforme DRAGEN n'effectue pas de tests visant des candidats au génotype diploïde. À la place, un allèle ALT est considéré comme étant un variant somatique candidat. Si le score SQ de tumeur > vc-sq-call-

threshold (la valeur par défaut est « 3 »), alors « 0/1 » est codé en dur dans le champ FORMAT/GT pour l'échantillon de tumeur et le champ FORMAT/AF produit une estimation de la fréquence allélique du variant somatique, soit entre [0,1].

- Si le score SQ de tumeur < vc-sq-call-threshold, le variant n'est pas produit dans le fichier VCF.
- Si le score SQ de tumeur > vc-sq-call-threshold, mais que le score SQ de tumeur < vc-sq-filter-threshold, le variant est produit dans le fichier VCF, mais la valeur du champ FILTER est « weak_evidence ».
- Si le score SQ de tumeur > vc-sq-call-threshold et que le score SQ de tumeur > vc-sq-filter-threshold, le variant est produit dans le fichier VCF et la valeur du champ FILTER est « weak_evidence ».
- La valeur par défaut de vc-sq-filter-threshold est « 17,5 » pour le mode relatif aux échantillons de tumeur et de référence et « 6,5 » pour le mode relatif aux échantillons de tumeur seulement.

L'exemple suivant montre un enregistrement VCF d'un échantillon somatique de tumeur et de référence. Si le score SQ de tumeur > vc-sq-call-threshold, mais que le score SQ de tumeur < vc-sq-filter-threshold, le champ FILTER est « weak_evidence ».

```
2 593701 . G A . weak_evidence
   DP=97;MQ=48.74;SQ=3.86;NLOD=9.83;FractionInformativeReads=1.000
   GT:SQ:AF:F1R2:F2R1:DP:SB:MB 0/0:9.83:33,0:0.000:14,0:19,0:33
   0/1:3.86:61,3:0.047:29,2:32,1:64:35,26,0,3:39,22,1,2
```

Analyse combinée multiéchantillons

La plateforme DRAGEN prend en charge l'analyse combinée multiéchantillons basée sur la généalogie et sur la population. Dans une analyse basée sur la généalogie, des échantillons provenant de la même espèce sont liés les uns par rapport aux autres, alors qu'ils ne sont pas liés dans une analyse basée sur la population.

Une analyse combinée nécessite un fichier gVCF pour chaque échantillon. Pour créer un fichier gVCF, exécutez la fonction de définition de petits variants en mode germinal en utilisant l'option *--vc-emit-ref-confidence gVCF*.

Le fichier gVCF contient des renseignements sur les positions des variants et sur les positions déterminées comme étant homozygotes par rapport au génome de référence. Pour les régions homozygotes, le fichier gVCF comprend des statistiques qui indiquent à quel point les lectures tolèrent l'absence de variants ou d'allèles alternatifs. Les séries de bases homozygotes contigües ayant des niveaux de confiance semblables sont regroupées en blocs, appelés blocs hom-ref. Les entrées du fichier gVCF ne sont pas toutes contigües. Une référence peut contenir des écarts non couverts par une ligne de variants ou un bloc hom-ref. Les écarts correspondent aux régions ne pouvant être définies. Une région est considérée comme ne pouvant être définie si aucune lecture correspondante à cette région n'a un score MAPQ supérieur à zéro. L'exemple suivant montre un fichier VCF combiné où un échantillon contient un variant et où les deux autres échantillons sont dans un écart de fichier gVCF. Les écarts sont représentés par « ./.:.: ».

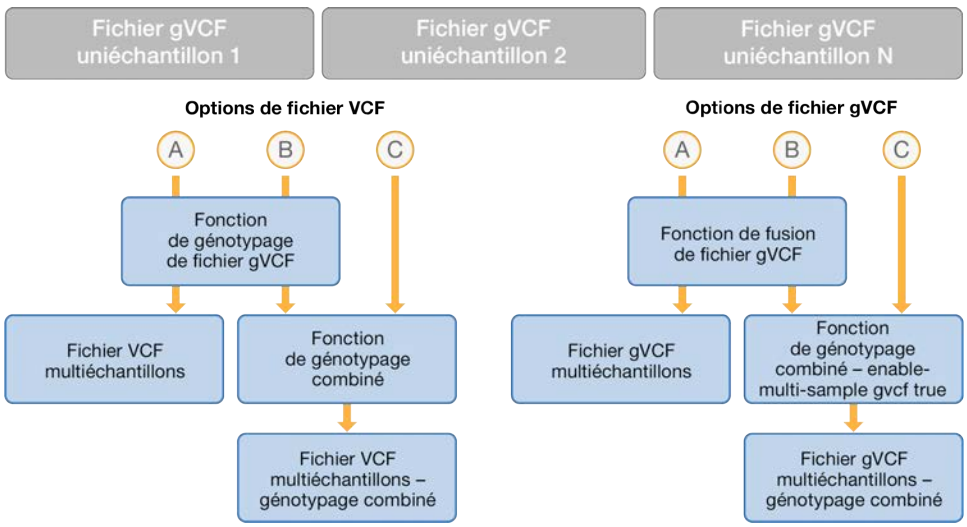
```
1 605262 . G A 13.41 DRAGENHardQUAL
   AC=2;AF=1.000;AN=2;DP=2;FS=0.000;MQ=14.00;QD=6.70;SOR=0.693
   GT:AD:AF:DP:GQ:FT:F1R2:F2R1:PL:GP ./.:.:.:.:.:LowDepth
   1/1:0,2:1.000:2:4:PASS:0,0:0,2:50,6,0:1.383e+01,4.943e+00,1.951e+00
   ./.:.:.:.:.:LowDepth
```

Mode population

La plateforme DRAGEN offre une option d'analyse basée sur une population pour effectuer une analyse combinée d'échantillons d'individus non apparentés. Pour activer le mode population, utilisez les fonctions de génotypage suivantes.

- **Fonction de génotypage de fichiers gVCF** – Utilise un ensemble de fichiers gVCF uniéchantillon comme entrée et produit un fichier VCF multiéchantillons qui contient une entrée pour tout variant qui se trouve dans un des fichiers gVCF. Les variants sont génotypés dans tous les échantillons en entrée à l'aide des renseignements des blocs de référence homozygote (hom-ref) si nécessaire. La fonction de génotypage de fichiers gVCF n'ajuste pas les génotypes en fonction des renseignements de la population. Consultez la section *Options de la fonction de génotypage de fichiers gVCF* à la page 50 pour obtenir des renseignements sur les options de ligne de commande disponibles.
- **Fonction de génotypage combiné** – Utilise des renseignements de l'ensemble de la cohorte pour améliorer l'exactitude des génotypes individuels. Vous pouvez utiliser en entrée un fichier VCF multiéchantillons, un fichier gVCF multiéchantillons ou un ensemble de fichiers gVCF uniéchantillon. Pour obtenir un fichier gVCF multiéchantillons en sortie, mettez à l'option --enable-multi-sample-gVCF la valeur « true » (activée). Consultez la section *Options de la fonction de génotypage combiné* à la page 51 pour obtenir des renseignements sur les options de ligne de commande disponibles.
- **Fonction de fusion de fichiers gVCF** – Utilise un fichier gVCF uniéchantillon et des fichiers gVCF multiéchantillons pour générer un fichier gVCF multiéchantillons. Les variants présents dans un échantillon en entrée sont génotypés dans tous les échantillons, un peu comme avec la fonction de génotypage de fichiers gVCF. La fonction de fusion de fichiers gVCF accepte toute combinaison de fichiers gVCF uniéchantillon et multiéchantillons en entrée. La génération du fichier de sortie pourrait être longue si beaucoup d'échantillons sont fusionnés. Pour la définition d'une très grande population, utilisez la fonction de génotypage de fichiers gVCF plutôt que la fonction de fusion de fichiers gVCF. Consultez la section *Options de la fonction de fusion de fichiers gVCF* à la page 51 pour obtenir des renseignements sur les options de ligne de commande disponibles.

La figure suivante illustre les différents trajets des flux de données entre la fonction de génotypage de fichiers gVCF, la fonction fusion de fichiers gVCF et la fonction de génotypage combiné.



Pour recevoir une liste de variants présents dans la cohorte et les génotypes de ce variant de chacun des membres de la cohorte, exécutez la fonction de génotypage de fichiers gVCF. Vous pouvez également exécuter la fonction de génotypage combiné pour créer un deuxième fichier VCF multiéchantillons. La sortie de la fonction de génotypage combiné affine les génotypes de l'échantillon selon les renseignements de la population. Si vous n'utilisez que la sortie de la fonction de génotypage de fichiers VCF, vous pouvez filtrer les variants peu fréquents afin de réduire le bruit si les variants ont une faible profondeur ou un génotype de mauvaise qualité. Utilisez le logiciel utilitaire libre bcftools sur le fichier de sortie pour filtrer les variants.

Mode généalogie

Utilisez le mode généalogie pour effectuer une analyse combinée des échantillons d'individus apparentés.

Pour utiliser le mode généalogie, mettez à l'option *--enable-joint-genotyping* la valeur « true » (activée) et utilisez l'option *--pedigree-file* pour spécifier le chemin d'accès vers le fichier de généalogie qui décrit les relations entre les panels.

Le fichier de généalogie doit être un fichier texte séparé par des tabulateurs avec une extension .ped. Les renseignements suivants sont requis :

En-tête de colonne	Description
Family_ID	L'identifiant de la généalogie.
Individual_ID	L'identifiant de l'individu.
Paternal_ID	L'identifiant du père de l'individu. S'il s'agit de la fondatrice, la valeur est « 0 ».
Maternal_ID	L'identifiant de la mère de l'individu. S'il s'agit de la fondatrice, la valeur est « 0 ».
Sex	Le sexe associé à l'échantillon. S'il s'agit du sexe masculin, la valeur est « 1 ». S'il s'agit du sexe féminin, la valeur est « 2 ».
Phénotype	Les données génétiques associées à l'échantillon. Si elles sont inconnues, la valeur est « 0 ». Si le problème génétique est absent, la valeur est « 1 ». Si le problème génétique est présent, la valeur est « 2 ».

Voici un exemple de fichier de généalogie d'entrée.

```
#Family_ID Individual_ID   Paternal_ID    Maternal_ID    Sex Phenotype
FAM001     NA12877_Father  0              0              1   1
FAM001     NA12878_Mother  0              0              2   1
FAM001     NA12882_Proband NA12877_Father NA12878_Mother 2   2
FAM001     NA12883_Proband NA12877_Father NA12878_Mother 1   0
```

Si vous faites une analyse combinée d'un trio, le proposant et ses deux parents, vous pouvez effectuer les analyses de variants du nombre de copies, de variants structurels et de petits variants structurels en parallèle grâce aux commandes suivantes.

1 En mode gVCF, effectuez une analyse uniéchantillon pour la définition de petits variants, des variants du nombre de copies et des variants structurels pour chaque échantillon avec une seule ligne de commande de la plateforme DRAGEN.

2 Effectuez une analyse de définition combinée *de novo* pour les petits variants, les variants du nombre de copies et les variants structurels.
La sortie produite est composée de trois fichiers *de novo* et VCF de sortie combinés. Pour plus de renseignements, consultez la section Définition combinée de novo.

Options de la fonction de génotypage

Cette section vous fournit des renseignements au sujet des options offertes pour chacune des fonctions de génotypage.

Options relatives au fichier gVCF

En plus des paramètres standards de l'étape de la fonction de définition de variants du logiciel hôte de la plateforme DRAGEN, les paramètres suivants peuvent être utilisés pour créer un fichier gVCF :

- *--vc-emit-ref-confidence*
  Pour activer la génération d'un fichier gVCF à bandes, mettez la valeur « GVCF ». Pour activer la génération d'un fichier gVCF de résolution de paires de bases, mettez la valeur « BP_RESOLUTION ». La génération d'un fichier gVCF à bandes est activée par défaut et est recommandée.
- *--vc-gvcf-gq-bands*
  L'option *--vc-gvcf-gq-bands* définit les valeurs des bandes de qualité du génotype. Si les valeurs de qualité du génotype pour les positions homozygotes à la référence se trouvent dans la même bande, les positions homozygotes à la référence sont regroupées dans un même bloc.
- *--vc-max-alternate-alleles*
  L'option *--vc-max-alternate-alleles* spécifie le nombre maximal d'allèles ALT produits dans le fichier de sortie VCF ou gVCF. La valeur par défaut est « 6 ».

Options de la fonction de génotypage de fichiers gVCF

La fonction de génotypage de fichiers gVCF utilise un ensemble de fichiers gVCF uniéchantillon pour produire un fichier VCF multiéchantillons qui contient une entrée pour tout variant qui se trouve dans un des fichiers gVCF. Les génotypes ne sont pas ajustés selon les renseignements de la population.

Les paramètres suivants peuvent être utilisés par la fonction de génotypage de fichiers gVCF :

- *--enable-gvcf-genotyper*
  Pour activer la fonction de génotypage de fichiers gVCF, mettez la valeur « true » (activée).
- *--ht-reference*
  Le fichier contenant la séquence de référence en format FASTA. L'option *--ht-reference* est obligatoire.
- *--output-directory*
  Le répertoire de sortie. L'option *--output-directory* est obligatoire.
- *--output-file-prefix*
  Le préfixe utilisé pour l'étiquetage de tous les fichiers de sortie. L'option *--output-file-prefix* est obligatoire.
- *--gg-output-format*
  Le format du fichier de sortie. La valeur par défaut est « vcf.gz ». Les formats pris en charge pour les fichiers de sortie sont vcf.gz, vcf et bcf. Seul le format vcf.gz est compatible avec la fonction de génotypage combinée. Si vous utilisez un autre format, vous pouvez convertir le fichier à l'aide de l'utilitaire libre bcftools.
- *--gg-regions*
  Le fichier spécifiant les régions dans lesquelles exécuter la fonction de génotypage de fichiers gVCF. Les variants qui se trouvent à l'extérieur de ces régions sont ignorés. Il peut s'agir d'un fichier BED ou d'une liste de régions génomiques spécifiées selon le modèle chromosome:start-end. Les régions génomiques peuvent être séparées par des virgules ou des sauts de lignes. Si vous utilisez des données d'exome ou d'enrichissement, spécifiez la liste des régions ciblées par les sondes afin de limiter le temps supplémentaire passé à traiter des variants de génotype non fiables se trouvant à l'extérieur des régions ciblées.
- *--gg-enable-concat*
  Concatène les résultats relatifs aux régions génomiques en un seul fichier de sortie. Par défaut, la valeur est « true » (activée).
- *--gg-max-alternate-alleles*
  Le nombre maximal d'allèles alternatifs. Par défaut, la valeur est « 50 ».

- *--gg-enable-indexing*
  Construit un index Tabix pour le fichier de sortie. Par défaut, la valeur est « false » (désactivée). La valeur de l'option *--gg-output-format* doit être « vcf.gz » pour pouvoir utiliser l'option *--gg-enable-indexing*.
- *--gg-drop-genotypes*
  Permet de produire seulement les résultats relatifs aux allèles pour chaque variant. Par défaut, la valeur est « false » (désactivée). L'option *--gg-drop-genotypes* équivaut à l'exécution de la fonction « view-G » de bcftools sur le fichier de sortie par défaut.
- *--gg-allele-list*
  [Facultatif] Force la production des génotypes pour les sites spécifiés. Le chemin d'accès vers un fichier vcf.gz ou bcf présentant les sites doit être compris.

Options de la fonction de génotypage combiné

Vous pouvez exécuter la fonction de génotypage combiné à partir d'un fichier VCF ou gVCF multiéchantillons ou directement à partir d'un ensemble de fichiers gVCF uniéchantillon.

Les paramètres suivants peuvent être utilisés par la fonction de génotypage combiné.

- *--enable-joint-genotyping*
  Pour exécuter la fonction de génotypage combiné, définissez la valeur à « true » (activée).
- *--output-directory*
  Le répertoire de sortie. L'option *--output-directory* est obligatoire.
- *--output-file-prefix*
  Le préfixe utilisé pour l'étiquetage de tous les fichiers de sortie. L'option *--output-file-prefix* est obligatoire.
- *-r*
  Le répertoire où se trouve la table de hachage.
- *--variant* ou *--variant-list*
  Spécifie le chemin d'accès vers un fichier gVCF. Vous pouvez spécifier plusieurs fichiers gVCF en utilisant plusieurs options *--variant*. Un maximum de 200 fichiers gVCF est pris en charge. Utilisez l'option *--variant-list* pour spécifier un fichier contenant une liste de fichiers gVCF à combiner en inscrivant un chemin d'accès vers un fichier de variants par ligne.
- *--pedigree-file*
  Spécifie le chemin d'accès vers un fichier de généalogie qui décrit les relations entre les échantillons. Pour en savoir plus, consultez la section *Mode généalogie* à la page 49.

Options de la fonction de fusion de fichiers gVCF

Vous pouvez exécuter la fonction de fusion de fichiers gVCF sur un ensemble de fichiers gVCF uniéchantillon et multiéchantillons pour créer un seul fichier de sortie gVCF. Pour la définition d'une très grande population, utilisez la fonction de génotypage de fichiers gVCF plutôt que la fonction de fusion de fichiers gVCF.

Les paramètres suivants sont disponibles pour la fonction de fusion de fichiers gVCF :

- *--enable-combinegvcfs*
  Pour exécuter la fonction de fusion de fichiers gVCF, mettez la valeur à « true » (activée).
- *--intermediate-results-dir*

[Facultatif] Permet de choisir un répertoire différent du répertoire de sortie pour les résultats intermédiaires, par exemple /staging/temp. Si vous exécutez la fonction de fusion de fichiers gVCF, l'option *--intermediate-results-dir* est recommandée. Pour en savoir plus, consultez la section *Emplacements des fichiers d'entrée et de sortie* à la page 5.

- *--output-directory*
  Le répertoire de sortie. L'option *--output-directory* est obligatoire.
- *--output-file-prefix*
  Le préfixe utilisé pour l'étiquetage de tous les fichiers de sortie. L'option *--output-file-prefix* est obligatoire.
- *-r*
  Le répertoire où se trouve la table de hachage.
- *--variant* ou *--variant-list*
  Spécifie le chemin d'accès vers un fichier gVCF. Vous pouvez spécifier plusieurs fichiers gVCF en utilisant plusieurs fois l'option *--variant*. Jusqu'à 200 fichiers gVCF peuvent être pris en charge, mais nous vous recommandons de ne pas en fusionner plus de 10. Utilisez l'option *--variant-list* pour spécifier un fichier contenant une liste de fichiers gVCF à combiner en inscrivant un chemin d'accès vers un fichier de variants par ligne.

Formats de sortie de l'analyse combinée

Deux fichiers de sortie de l'analyse combinée sont disponibles :

- **Fichier VCF multiéchantillons** – Un fichier VCF avec une colonne qui contient les renseignements de génotype pour chacun des échantillons d'entrée selon les variants d'entrée.
- **Fichier gVCF multiéchantillons** – Un fichier gVCF qui enrichit le contenu d'un fichier VCF multiéchantillons, tout comme un fichier gVCF enrichit un fichier VCF uniéchantillon. Entre les sites de variants, le fichier gVCF multiéchantillons contient des statistiques qui décrivent le niveau de confiance quant à l'homozygotie de chaque échantillon par rapport au génome de référence. Le format gVCF multiéchantillons est pratique, car il combine les résultats d'une généalogie ou d'une petite cohorte dans un seul fichier. Si vous utilisez un grand nombre d'échantillons, les variations dans la couverture ou dans un des échantillons d'entrée créent un nouveau bloc de référence homozygote (hom-ref), ce qui entraîne une fragmentation importante de la structure de blocs et un fichier de sortie volumineux qui peut être long à créer.

Champs FORMAT des blocs de référence homozygote (hom-ref)

Dans les blocs hom-ref, les champs FORMAT suivants sont calculés séparément.

- **FORMAT/DP** – Les valeurs représentent la profondeur minimale pour l'ensemble des positions sur la bande.
- **FORMAT/AD** – Les valeurs représentent la position de la valeur correspondant à la profondeur médiane sur la bande.
- **FORMAT/AF** – Les valeurs sont basées sur celles du champ FORMAT/AD.
- **FORMAT/PL** – Les valeurs représentent les probabilités, selon l'algorithme Phred, pour chaque échantillon selon l'hypothèse « par génotype ». Pour une position de variant précis dans un trio d'échantillons, le score de confiance qu'il s'agisse d'un variant *de novo* chez le proposant est calculé à partir des valeurs de probabilités Phred et des probabilités précédentes de chaque échantillon. Dans les

cas de polymorphisme mononucléotidique, les valeurs de probabilités Phred sont reportées dans le champ FORMAT/SPL. Pour les variants INDEL, la valeur de probabilité Phred est calculée à une étape de définition de la généalogie commune à partir du champ FORMAT/ICNT reporté dans le fichier gVCF.

- FORMAT/SPL et FORMAT/ICNT – Les paramètres reportés dans les enregistrements de fichier gVCF, y compris les enregistrements hom-ref et les enregistrements de variants. FORMAT/SPL et FORMAT/ICNT sont tous les deux composés de deux valeurs. La première valeur correspond au nombre de lectures pour lesquelles aucun indel ne se trouve à la position et la deuxième valeur au nombre de lectures pour lesquelles un indel s'y trouve. Chaque valeur du champ FORMAT/ICNT représente le maximum de la valeur pour toutes les positions sur la bande. Chaque valeur du champ FORMAT/SPL représente le minimum.

Dans l'exemple de bloc hom-ref suivant, la valeur ICNT indique si chaque échantillon contient un indel à la position d'intérêt. Si le proposant contient un indel à cette position et que la valeur ICNT des parents n'indique aucune lecture à l'appui de l'indel, alors le score de confiance qu'il s'agisse d'un indel *de novo* est élevé.

```
chr1 10288 . C <NON_REF> . PASS END=10290 GT:AD:DP:GQ:MIN_DP:PL:SPL:ICNT
   0/0:131,4:135:69:132:0,69,1035:0,125,255:23,1 chr1 10291 . C T,<NON_
   REF> 38.45 PASS
   DP=100;MQ=24.72;MQRankSum=0.733;ReadPosRankSum=4.112;FractionInformativ
   eReads=0.600;R2_5P_bias=0.000
   GT:AD:AF:DP:F1R2:F2R1:GQ:PL:SPL:ICNT:GP:PRI:SB:MB
   0/1:28,32,0:0.533,0.000:60:20,21,0:8,11,0:15:73,0,12,307,157,464:255,0,
   255:23,10:3.8452e+01,1.3151e-
   01,1.5275e+01,3.0757e+02,1.9173e+02,4.5000e+02:0.00,34.77,37.77,34.77,6
   9.54,37.77:4,24,7,25:8,20,14,18
```

La valeur ICNT est propre à la plateforme DRAGEN. La fonction de définition de variants de GATK ne produit pas de valeur ICNT.

Filtrage des variants de novo

L'étape du filtrage vise à identifier les définitions de variants de novo du flux de travail de définition combinée dans les régions présentant des modifications de ploïdie. Puisque la définition de novo peut réduire la spécificité dans les régions où au moins un des membres de la généalogie présente des génotypes non diploïdes, le filtrage de variants de novo marque les variants pertinents et permet donc d'améliorer la spécificité de l'ensemble de la définition.

Selon la définition de variants structurels et du nombre de copies de la généalogie, la valeur du champ FORMAT/DN du proposant passe de « DeNovo » à « DeNovoSV » ou à « DeNovoCNV » si le variant de novo chevauche un variant structurel ou un VNC dont la ploïdie est modifiée, respectivement. Tous les autres détails sur les variants demeurent inchangés et tous les variants du fichier VCF d'entrée sont également présents dans le fichier VCF filtré de sortie. Les variants structurels ou les VNC qui ne présentent pas de modification de la ploïdie, par exemple dans le cas d'inversions, ne sont pas pris en considération dans le filtrage. Voici un exemple de définition de variant structurel dans le fichier VCF d'entrée :

```
chr1 234710899 . T C 44.74 PASS
AC=1;AF=0.167;AN=6;DP=73;FS=4.720;MQ=250.00;MQRankSum=5.310;QD=1.15;Rea
dPosRankSum=1.366;SOR=0.251
GT:AD:AF:DP:GQ:FT:F1R2:F2R1:PL:GL:GP:PP:DQ:DN
0/1:21,18:0.462:39:48:PASS:14,10:7,8:84,0,50:-8.427,0,-
```

```
5:4.950e+01,7.041e-05,5.300e+01:15,0,120:3.2280e-01:DeNovo
0/0:13,0:0.000:11:30:PASS:.:.:0,30,450:.:.:10,0,227
0/0:25,0:0.000:22:60:PASS:.:.:0,60,899:.:.:0,33,227
```

Le chevauchement d'une répétition du variant structurel chez le proposant, la mère ou le père est représenté de la manière suivante dans le fichier VCF filtré de sortie :

```
chr1 234710899 . T C 44.74 PASS
AC=1;AF=0.167;AN=6;DP=73;FS=4.720;MQ=250.00;MQRankSum=5.310;QD=1.15;Rea
dPosRankSum=1.366;SOR=0.251
GT:AD:AF:DP:GQ:FT:F1R2:F2R1:PL:GL:GP:PP:DQ:DN
0/1:21,18:0.462:39:48:PASS:14,10:7,8:84,0,50:-8.427,0,-
5:4.950e+01,7.041e-05,5.300e+01:15,0,120:3.2280e-01:DeNovoSV
0/0:13,0:0.000:11:30:PASS:.:.:0,30,450:.:.:10,0,227
0/0:25,0:0.000:22:60:PASS:.:.:0,60,899:.:.:0,33,227
```

L'exemple suivant montre une ligne de commande permettant d'exécuter le filtrage de novo à partir des fichiers retournés par les flux de travail de définition combinée :

```
dragen \
   --dn-enable-denovo-filtering true \
   --dn-input-joint-vcf <JOINT_SMALL_VARIANT_VCF> \
   --dn-output-joint-vcf <OUTPUT_VCF> \
   --dn-sv-vcf <JOINT_SV_VCF> \
   --dn-cnv-vcf <JOINT_CNV_VCF> \
   --enable-map-align false
```

Options de filtres pour variants de novo

Les options suivantes sont utilisées pour filtrer les variants de novo :

- *--dn-input-vcf* – Le fichier VCF de petits variants combiné à filtrer après l'étape de définition de novo.
- *--dn-output-vcf* – L'emplacement d'écriture du fichier VFC filtré. Si l'emplacement n'est pas spécifié, le fichier VCF d'entrée est écrasé.
- *--dn-sv-vcf* – Le fichier VCF de variants structurels combiné provenant de l'étape de définition de variants structurels. En cas d'omission, les vérifications avec les variants structurels qui se chevauchent sont sautées.
- *--dn-cnv-vcf* – Le fichier VCF de variants structurels combiné provenant de l'étape de définition de variant du nombre de copies (VNC). En cas d'omission, les vérifications avec les VNC qui se chevauchent sont sautées.

Filtres des variants germinaux

La plateforme DRAGEN fournit des filtres de variants postproduction des fichiers VCF selon les annotations présentes dans les enregistrements VCF. Les filtres de variants par défaut et autres sont décrits ci-dessous. Toutefois, en raison de la nature des algorithmes de la plateforme DRAGEN, qui incorpore l'hypothèse d'erreurs corrélées à partir du cœur de la fonction de définition des variants, le pipeline présente des capacités accrues de distinction des véritables variants du bruit. Par conséquent, la dépendance aux filtres post-production des fichiers VCF est significativement réduite. Pour cette raison, les filtres post-production des fichiers VCF par défaut de la plateforme DRAGEN sont très simples.

Filtres de variants par défaut

Les filtres par défaut du pipeline germinal sont les suivants :

- ##FILTER=<ID=DRAGENHardQUAL,Description="Set if true:QUAL < 10.4139">
- ##FILTER=<ID=PloidyConflict,Description="Genotype call from variant caller not consistent with chromosome ploidy">
- ##FILTER=<ID=lod_fstar,Description="Variant does not meet likelihood threshold (default threshold is 6.3)">
- DRAGENHardQUAL : Pour tous les contigs autres que le contig mitochondrial, le filtre par défaut consiste en l'imposition d'un seuil pour la valeur QUAL (qualité du variant) seulement.
- lod_fstar : Pour le contig mitochondrial, le filtre par défaut consiste en l'imposition d'un seuil pour le score LOD (limite de détection) seulement.
- PloidyConflict : Ce filtre est appliqué à toutes les définitions de variants d'un chrY sur un sujet féminin, si « female » est spécifié dans la ligne de commande.

Autres filtres de variants

La plateforme DRAGEN prend en charge le filtre de base des définitions de variants, comme décrit dans le fichier VCF standard. Vous pouvez appliquer autant de filtres que vous voulez grâce à l'option *--vc-hard-filter*, qui vous permet de créer une liste d'expressions séparée par des points-virgules, comme suit :

<filter ID>:<snp|indel|all>:<list of criteria>,

La liste de critères est une liste d'expressions délimitées par l'opérateur || (OU) dans ce format :

<annotation ID> <comparison operator> <value>

La signification de ces expressions est la suivante :

- **filterID** – Le nom du filtre, qui est entré dans la colonne FILTER du fichier VCF pour les définitions filtrées grâce à cette expression.
- **snp/indel/all** – Le sous-ensemble de définitions de variants auquel l'expression doit être appliquée.
- **annotation ID** – L'annotation de l'enregistrement de la définition de variant pour laquelle les valeurs devraient être filtrées. Les annotations prises en charge comprennent « FS », « MQ », « MQRankSum », « QD » et « ReadPosRankSum ».
- **comparison operator** – L'opérateur de comparaison numérique à utiliser pour comparer la valeur de filtre spécifiée. Les opérateurs pris en charge comprennent « < », « ≤ », « = », « ≠ », « ≥ » et « > ».

Par exemple, l'expression suivante apposerait l'étiquette « SNP filter » à tout polymorphisme mononucléotidique dont « FS < 2,1 » ou « MQ < 100 » et l'étiquette « indel filter » à tout enregistrement dont « FS < 2,2 » ou « MQ < 110 » :

```
--vc-hard-filter="SNP filter:snp:FS < 2.1 || MQ < 100; indel
   filter:indel:FS < 2.2 || MQ < 110"
```

Cet exemple sert à des fins d'illustration seulement et ne devrait PAS être utilisé dans un fichier de sortie de la plateforme DRAGEN V3. Illumina recommande plutôt l'utilisation des filtres par défaut.

La seule opération prise en charge pour combiner des comparaisons de valeurs est « OR ».
Les combinaisons arithmétiques de plusieurs annotations ne sont pas prises en charge. Des expressions plus complexes pourraient être prises en charge dans des versions ultérieures.

Filtre de biais de l'orientation

Le filtre de biais de l'orientation est conçu pour réduire le bruit qui est typiquement associé aux éléments suivants :

- Les artéfacts préadaptation introduits durant la préparation de la librairie génomique (p. ex., une combinaison contaminants causés par la chaleur, le cisaillement et les métaux pourrait entraîner des paires de bases 8-oxoguanine et cytosine ou adénine, menant ultimement à des mutations par transversion G→T durant l'amplification PCR), ou
- Les artéfacts fixés au formol et imprégnés à la paraffine (FFPE). Les artéfacts FFPE découlent de la désamination des cytosines, ce qui se traduit en mutations par transition C→T.

Le filtre de biais de l'orientation ne peut être utilisé que dans les pipelines en mode somatique. Pour activer le filtre, mettez à l'option *--vc-enable-orientation-bias-filter* la valeur « true » (activée). La valeur par défaut est « false » (désactivée).

Le type d'artéfacts à filtrer peut être spécifié grâce à l'option *--vc-orientation-bias-filter-artifacts*. Les valeurs par défaut sont C/T,G/T, qui correspondent aux artéfacts OxoG et FFPE. Les valeurs valides sont C/T ou G/T ou C/T,G/T,C/A.

Un artéfact (ou un artéfact et son complément inverse) ne peut pas apparaître plus d'une fois dans la liste. Par exemple, la combinaison des valeurs C/T,G/A n'est pas valide, puisque C→G et T→A sont des compléments inverses.

Annotation dbSNP

En mode germinal ou somatique (échantillons de tumeur et de référence ou échantillon de tumeur seulement), la plateforme DRAGEN peut rechercher des définitions de variants dans une base de données dbSNP et ajouter des annotations pour toute correspondance trouvée. Pour activer la recherche dans la base de données dbSNP, mettez à l'option *--dbsnp* le chemin d'accès vers le fichier VCF ou .vcf.gz de la base de données dbSNP, qui doit être triée selon la référence.

Pour chaque définition de variant dans le fichier VCF de sortie, si la définition correspond à une entrée de la base de données pour les valeurs CHROM, POS, REF et au moins une ALT, alors l'identifiant rsID de l'entrée est copié dans la colonne ID du fichier VCF de sortie pour cette définition. De plus, la plateforme DRAGEN ajoute une annotation DB (base de données) dans le champ INFO pour les définitions trouvées dans la base de données.

La plateforme DRAGEN met les définitions de variants en correspondance selon le nom de la séquence ou du contig de référence. Il n'y a toutefois pas d'autre manière de vérifier que la référence utilisée pour construire la base de données dbSNP est la même que celle utilisée par la référence pour l'alignement et la définition de variants. Assurez-vous que les contigs qui se trouvent dans la base de données de l'annotation sélectionnée correspondent à ceux qui se trouvent dans la référence de l'alignement et de la définition de variants.

Fichier VCF du panel de référence

Lorsque la plateforme DRAGEN est utilisée en mode somatique (échantillons de tumeur et de référence ou échantillon de tumeur seulement, elle recherche des définitions de variants dans un fichier VCF de panel de référence. Le fichier VCF de panel de référence doit être préalablement généré et représente un ensemble de variants détectés par le pipeline somatique de la plateforme DRAGEN lorsqu'il est exécuté sur un ensemble d'échantillons de référence (qui ne sont pas nécessairement mis en correspondance avec le sujet duquel provient l'échantillon de tumeur). Idéalement, toutefois, le fichier VCF du panel de référence devrait

provenir d'échantillons de référence prélevés par le même instrument de préparation de librairies ou de séquençage, de manière à ce que les erreurs systémiques pouvant survenir durant le séquençage ou la préparation de librairies se reflètent dans le fichier VCF de panel de référence.

- *--panel-of-normals*

  Spécifie un fichier VCF de panel de référence. Lorsqu'un fichier VCF de panel de référence est utilisé comme fichier d'entrée, si un variant somatique est trouvé dans au moins un échantillon du fichier (qui peut contenir plusieurs douzaines d'échantillons), il est marqué comme « panel_of_normals » (panel de référence) dans la colonne FILTER du fichier VCF de sortie.

Génération automatique d'un fichier MD5SUM pour les fichiers VCF

Un fichier MD5SUM est généré automatiquement pour les fichiers VCF de sortie. Ce fichier se trouve dans le même répertoire de sortie et a le même nom que le fichier VCF de sortie, mais auquel est ajouté une extension .md5sum. Par exemple, whole_genome_run_123.vcf.md5sum. Le fichier MD5SUM est un fichier à une seule ligne qui contient la fonction md5sum du fichier VCF de sortie. Cette fonction md5sum correspond exactement au résultat de la commande md5sum de Linux .

Définition de variants du nombre de copies

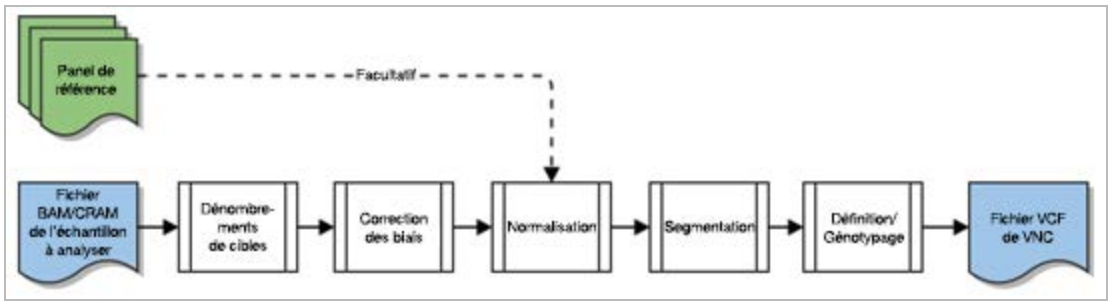
Le pipeline de variants du nombre de copies (VNC) de la plateforme DRAGEN peut définir des événements relatifs aux VNC à partir de données de séquençage nouvelle génération. Ce pipeline prend en charge plusieurs applications dans une seule interface à l'aide du logiciel hôte de la plateforme DRAGEN, y compris le traitement de données de séquençage de génomes entiers et d'exomes entiers aux fins d'analyse germinale.

Le pipeline de VNC de la plateforme DRAGEN prend en charge deux modes fonctionnement pour la normalisation. Les deux modes appliquent différentes techniques de normalisation pour traiter des biais variables selon l'application, par exemple le séquençage de génomes entiers et d'exomes entiers. Bien que les paramètres d'options par défaut tentent de fournir le meilleur équilibre entre la vitesse et l'exactitude, un flux de travail particulier peut nécessiter des paramètres d'option plus adaptés.

Flux de travail de VNC

Le pipeline de VNC de la plateforme DRAGEN suit le flux de travail montré dans la figure suivante.

Figure 3 Flux de travail du pipeline de VNC de la plateforme DRAGEN



Ce pipeline utilise plusieurs aspects de la plateforme DRAGEN qui sont disponibles dans d'autres pipelines, comme l'accélération matérielle et le traitement efficace du système. Pour activer le traitement des VNC dans le logiciel hôte de la plateforme DRAGEN, mettez à l'option de ligne de commande *--enable-cnv* la valeur « true » (activée).

Le pipeline de VNC comprend les modules de traitement suivants :

- Target Counts (Dénombrements de cibles) – La compartimentation des dénombrements de lectures et des autres signaux à partir des alignements.
- Bias Correction (Correction des biais) – La correction des biais intrinsèques aux systèmes.
- Normalization (Normalisation) – La détection des niveaux de ploïdie de référence et la normalisation de l'échantillon à analyser.
- Segmentation – La détection des points de rupture en passant par la segmentation du signal normalisé.
- Calling/Genotyping (Définition/Génotypage) – La définition des seuils, des scores, des qualifications et des filtres des événements putatifs comme VNC.

Le module de normalisation peut aussi accepter un panel de référence, qui est utilisé lorsqu'une cohorte ou des échantillons de population sont faciles à obtenir. Tous les autres modules utilisent les différents algorithmes relatifs aux VNC.

Analyse du flux relatif au signal

Les figures qui suivent montrent un aperçu général des étapes du pipeline de VNC de la plateforme DRAGEN alors que le signal passe par les différentes étapes. Ces figures sont des exemples et ne sont pas identiques aux graphiques générés par le pipeline de VNC de la plateforme DRAGEN.

La première étape du pipeline VNC de la plateforme DRAGEN est celle des dénombrements de cibles où les signaux comme le dénombrement de lectures et les paires mal appariées sont extraits et placés dans des intervalles de cibles.

Figure 4 Signal du dénombrement de lectures

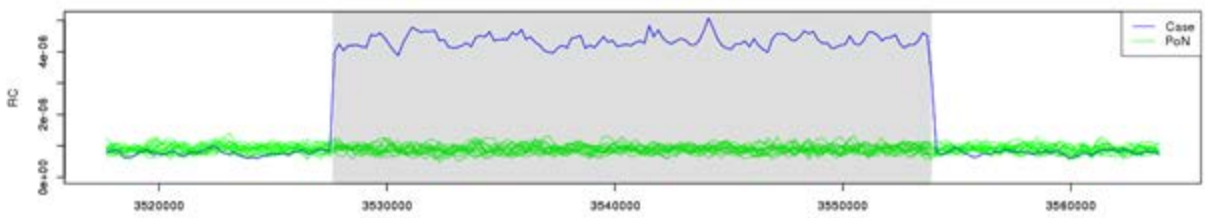
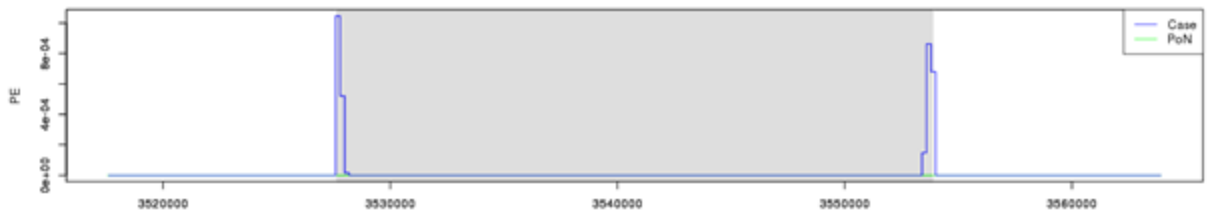
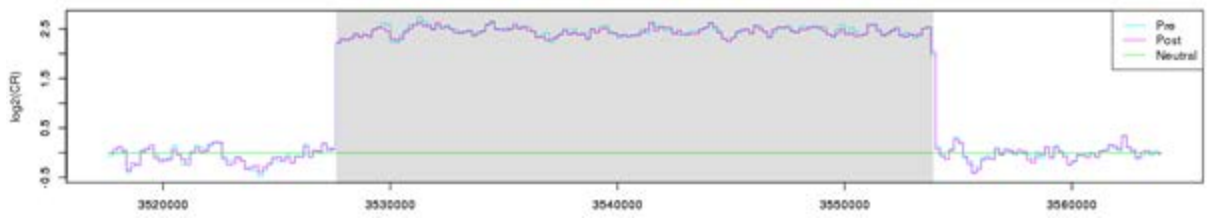


Figure 5 Signal de paires mal appariées



Ensuite, l'échantillon à analyser est normalisé par rapport au panel de référence ou à l'estimation du niveau de ploïdie de référence et tous les autres biais sont soustraits du signal pour amplifier tous les signaux d'événement.

Figure 6 Avant et après la normalisation de la moyenne



Le signal normalisé est alors segmenté à l'aide d'un des algorithmes de segmentation disponibles, puis les événements sont définis à partir des segments.

Figure 7 Segments

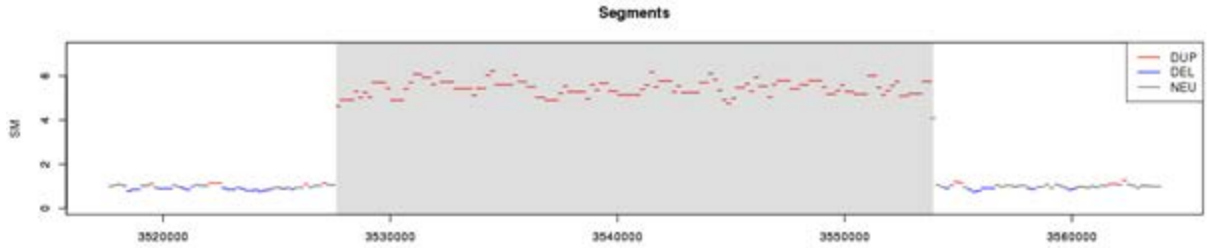


Figure 8 Événements définis

Les événements sont ensuite notés et produits dans le fichier VCF de sortie.

Options du pipeline de VNC

Les options suivantes sont les principales options communiquées au logiciel hôte de la plateforme DRAGEN pour contrôler le pipeline de VNC. Le fichier d'entrée peut être un fichier BAM ou CRAM. Si vous utilisez la fonction de cartographie et d'alignement de la plateforme DRAGEN, des fichiers FASTQ peuvent être utilisés.

- *--bam-input* – Le fichier BAM à traiter.
- *--cram-input* – Le fichier CRAM à traiter.
- *--enable-cnv* – Permet d'activer ou de désactiver le traitement de VNC. La valeur « true » (activée) active le traitement de VNC.
- *--enable-map-align* – Active le module de la fonction de cartographie et d'alignement. La valeur par défaut est « true » (activée) afin que toutes les lectures d'entrée soient cartographiées et alignées.
- *--fastq-file1*, *--fastq-file2* – Le fichier ou fichiers FASTQ à traiter.
- *--output-directory* – Le répertoire de sortie pour le stockage des résultats.
- *--output-file-prefix* – Le préfixe du fichier de sortie à appliquer à tous les noms de fichiers de résultats.

- *--ref-dir* – Le répertoire de la table de hachage du génome de référence de la plateforme DRAGEN.

Fichiers d'entrée du pipeline de variants du nombre de copies

Le pipeline de VNC de la plateforme DRAGEN prend en charge plusieurs formats d'entrée. Le format le plus courant est un fichier BAM ou CRAM déjà cartographié et aligné. Si vos données n'ont pas encore été cartographiées et alignées, consultez la section *Générer un fichier d'alignement* à la page 60.

Pour lancer le pipeline de VNC de la plateforme DRAGEN directement avec un fichier d'entrée FASTQ sans générer de fichier BAM ou CRAM, consultez la section *Transmission d'alignements* à la page 61), qui présente les étapes de transmission d'enregistrements d'alignements directement à partir de l'étape de cartographie et d'alignement de la plateforme DRAGEN.

Table de hachage de référence

Dans le cas du pipeline de VNC de la plateforme DRAGEN, la table de hachage doit être générée en mettant à l'option *--enable-cnv* la valeur « true » (activée) en plus de toutes les autres options requises par les autres pipelines. Lorsque la valeur de l'option *--enable-cnv* est « true » (activée), la commande *dragen* génère une nouvelle carte d'unicité de k-mers pour contrer les biais de cartographiabilité. Le fichier de carte d'unicité de k-mers n'a besoin d'être généré qu'une fois par table de hachage de référence. Cela prend environ 1,5 heure pour chaque génome humain entier.

La table de hachage de référence est une représentation binaire prégénérée du génome de référence. Pour en savoir plus sur la génération d'une table de hachage, consultez la section *Préparation d'un génome de référence* à la page 141.

L'exemple suivant montre une ligne de commande générant une table de hachage.

```
dragen \
   --build-hash-table true \
   --ht-reference <FASTA> \
   --output-directory <OUTPUT> \
   --enable-cnv true \
   --enable-rna true
```

Générer un fichier d'alignement

La ligne de commande suivante montre comment exécuter le pipeline de cartographie et d'alignement de la plateforme DRAGEN selon votre type de fichier d'entrée. Le pipeline génère un fichier d'alignement au format BAM ou CRAM qui peut ensuite être utilisé dans le pipeline de VNC de la plateforme.

Vous devez générer des fichiers d'alignement pour tous les échantillons qui n'ont pas déjà été cartographiés et alignés, y compris les alignements à utiliser comme référence à l'étape de la normalisation. Chaque échantillon doit avoir un identifiant unique qui est spécifié à l'aide de l'option *--RGSM*. Pour les fichiers d'entrée BAM et CRAM, l'identifiant de l'échantillon est tiré du fichier, l'option *--RGSM* n'est donc pas nécessaire.

Voici un exemple d'une commande permettant de cartographier et d'aligner un fichier FASTQ :

```
dragen \
   -r <HASHTABLE> \
   -1 <FASTQ1> \
   -2 <FASTQ2> \
   --RGSM <SAMPLE> \
   --RGID <RGID> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
   --enable-map-align true
```

Voici un exemple d'une commande permettant de cartographier et d'aligner un fichier BAM existant :

```
dragen \
   -r <HASHTABLE> \
   --bam-input <BAM> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
   --enable-map-align true
```

Voici un exemple d'une commande permettant de cartographier et d'aligner un fichier CRAM existant :

```
dragen \
   -r <HASHTABLE> \
   --cram-input <CRAM> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
   --enable-map-align true
```

Transmission d'alignements

La plateforme DRAGEN peut cartographier et aligner des échantillons FASTQ, puis les transmettre directement aux fonctions de définition en aval, par exemple à la fonction de définition de VNC et à la fonction de définition de variants haplotypes. Ce processus vous permet de sauter la génération d'un fichier BAM ou CRAM, donc de contourner le besoin de stocker des fichiers supplémentaires.

Pour transmettre les alignements directement dans le pipeline de VNC, l'échantillon FASTQ doit être analysé selon un flux de travail normal de cartographie et d'alignement par la plateforme DRAGEN, puis il est possible d'ajouter des arguments supplémentaires pour activer les VNC. L'exemple suivant montre une ligne de commande permettant de cartographier et d'aligner un fichier FASTQ avant de l'envoyer au pipeline de VNC.

```
dragen \
   -r <HASHTABLE> \
   -1 <FASTQ1> \
   -2 <FASTQ2> \
   --RGSM <SAMPLE> \
   --RGID <RGID> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
   --enable-map-align true \
   --enable-cnv true \
   --cnv-enable-self-normalization true
```

Pour en savoir plus au sujet de l'analyse des VNC avec l'utilisation simultanée de la fonction de définition de variants haplotypes, consultez la section *Définition simultanée de variants du nombre de copies et de petits variants* à la page 74.

Dénombrements de cibles

L'étape de dénombrements de cibles est la première étape de traitement du pipeline de VNC de la plateforme DRAGEN. Cette étape compartimente les alignements par intervalles. Le format principal d'analyse des VNC est le fichier de dénombrements de cibles, qui contient les signaux extraits des alignements qui doivent être utilisés dans le traitement en aval. La stratégie de compartimentation, la taille des intervalles et leurs limites sont contrôlées par les options de génération du nombre de cibles et la technique de normalisation utilisée.

Lorsque vous travaillez avec les données d'une séquence de génome entier, les intervalles sont autogénérés à partir de la table de hachage de référence. Seuls les contigs principaux de la table de hachage de référence sont pris en considération pour la compartimentation. Vous pouvez spécifier des contigs à sauter grâce à l'option *--cnv-skip-contig-list*.

Quand vous travaillez avec les données d'une séquence d'exome entier, le fichier BED cible fournit avec l'option *--cnv-target-bed* est utilisé pour déterminer les intervalles d'analyse.

L'étape de dénombrements de cibles génère un fichier .target.counts, qui peut ensuite être utilisé au lieu d'un fichier BAM ou CRAM si l'option *--cnv-input* est spéficiée à l'étape de la normalisation. Le fichier .target.counts est un fichier intermédiaire pour le pipeline de VNC de la plateforme DRAGEN. Il ne devrait pas être modifié.

Le fichier .target.counts est un fichier texte délimité par des tabulations. Il contient les colonnes suivantes :

- Contig identifier (Identifiant de contig)
- Start Position (Position de départ)
- End position (Position finale)
- Target interval name (Nom de l'intervalle de cibles)
- Count of alignments in this interval (Dénombrement d'alignements dans cet intervalle)
- Count of improperly paired alignments in this interval (Dénombrement d'alignements mal appariés dans cet intervalle)

Voici un exemple d'un fichier *.target.counts :

```
contig start stop name SampleName improper_pairs
1 565480 565959 target-wgs-1-565480 7 6
1 566837 567182 target-wgs-1-566837 9 0
1 713984 714455 target-wgs-1-713984 34 4
1 721116 721593 target-wgs-1-721116 47 1
1 724219 724547 target-wgs-1-724219 24 21
1 725166 725544 target-wgs-1-725166 43 12
1 726381 726817 target-wgs-1-726381 47 14
1 753243 753655 target-wgs-1-753243 31 2
1 754322 754594 target-wgs-1-754322 27 0
1 754594 755052 target-wgs-1-754594 41 0
```

Génome entier

Si des échantillons de génome entier sont utilisés, alors la largeur réelle des intervalles de cibles est spécifiée à l'aide de l'option *--cnv-interval-width*. Plus la couverture de l'échantillon est grande, plus la détection peut se faire avec une résolution élevée. Cette option est importante quand vous analysez un panel de référence, car tous les échantillons doivent avoir des intervalles identiques. Pour la normalisation automatique, la largeur réelle peut être plus grande que la valeur spécifiée.

Couverture de séquençage du génome entier	Résolution recommandée*
5x	1000 pb
10x	1000 pb
20x	500 pb
30x	250 pb
50x	250 pb

* Vous pouvez choisir de réduire la résolution pour accélérer le traitement, ou le contraire.

Les intervalles sont générés automatiquement pour chaque contig de la référence. Vous pouvez spécifier une liste de contigs à sauter avec l'option *--cnv-skip-contig-list*. Cette option prend comme paramètre une liste d'identifiants de contig séparés par des virgules. Les identifiants de contig doivent correspondre aux valeurs de la table de hachage de référence utilisée. Par défaut, seuls les chromosomes mitochondriaux sont sautés. Les contigs non primaires ne sont jamais traités.

Par exemple, pour sauter les chromosomes M, X et Y, utilisez l'option suivante :

```
--cnv-skip-contig-list "chrM,chrX,chrY"
```

Exome entier

Si des échantillons d'exome entier sont utilisés, un fichier BED de cibles doit être fourni à l'aide de l'option *--cnv-target-bed $TARGET_BED*.

Le fichier BED de cibles doit avoir une ligne d'en-tête et une quatrième colonne qui contient le nom de la cible. Voici une simple commande awk qui transforme un fichier BED de cibles en entrée (sans en-tête ou lignes de commentaires) en un fichier adapté au variants du nombre de copies.

```
cat <(echo -e "contig\tstart\tstop\tname") \
<(awk '{print $1"\t"$2"\t"$3"\ttarget-"NR}' $ORIGINAL_BED)
```

Un fichier BED ordinaire sans en-tête peut également être utilisé. Dans ce cas, les noms de cible de la quatrième colonne sont générés automatiquement par l'algorithme de variants du nombre de copies lors de l'étape des dénombrements de cibles. Pour utiliser un fichier BED standard, assurez-vous que celui-ci n'a pas d'en-tête. Dans ce cas, toutes les colonnes après la troisième sont ignorées, comme avec la fonction de définition de variants de la plateforme DRAGEN.

Options relatives aux dénombrements de cibles

Les options suivantes permettent de contrôler la génération de dénombrements de cibles.

- *--cnv-counts-method* – Spécifie la méthode de dénombrement pour qu'un alignement soit dénombré dans un compartiment de cibles. Les valeurs peuvent être « midpoint » (milieu), « start » (début) ou « overlap » (chevauchement). La valeur par défaut est « overlap » quand vous utilisez l'approche par

panel de référence, ce qui signifie que si un alignement chevauche une partie du compartiment de cibles, il est dénombré dans ce compartiment. En mode de normalisation automatique, la valeur par défaut de la méthode de dénombrement est « start ».

- *--cnv-min-mapq* – Spécifie la valeur minimale de MAPQ nécessaire pour qu'une lecture soit dénombrée lors de la génération des dénombrements de cibles. Veuillez prendre note que les lectures avec une valeur de MAPQ de 0 sont toujours dénombrées, peu importe la valeur de ce paramètre. La valeur par défaut est « 20 ».
- *--cnv-target-bed* – Spécifie le fichier adéquatement formaté (au format BED) précisant l'intervalle de cibles couvert par l'échantillon. À utiliser pour les analyses de séquençage d'exome entier.
- *--cnv-interval-width* – Spécifie la largeur de l'intervalle d'échantillonnage pour le séquençage du génome entier de variants du nombre de copies. Cette option permet de régler la taille réelle de l'intervalle. La valeur par défaut est « 1000 ».
- *--cnv-skip-contig-list* – Spécifie une liste à valeurs séparées par des virgules d'identifiants de contigs à sauter lors de la génération d'intervalles pour l'analyse de séquençage du génome entier. Par défaut, les contigs qui sont sautés, s'ils ne sont pas spécifiés, sont « chrM,MT,m,chrm ».

Régions éliminées des dénombrements de cibles

Pour le séquençage de génome entier, quand un fichier BED n'est pas spécifié pour une référence donnée, les mêmes intervalles devraient être générés chaque fois. Les intervalles créés prennent en considération la cartographiabilité du génome de référence à l'aide d'une cartographie d'unicité des k-mers créée lors de la construction de la table de hachage. Une région éliminée est une région complexe où ne sont pas dénombrés les alignements, ce qui entraîne un intervalle manquant dans l'analyse. Les régions éliminées comprennent les centromères, les télomères et les régions à faible complexité. S'il y a suffisamment de signal dans les régions flanquantes, un événement peut tout de même englober ces régions éliminées et ce, même si le dénombrement d'alignements n'est pas effectué dans ces régions. L'événement est géré par l'étape de segmentation.

Correction des biais GC

Les biais GC mesurent la relation entre le contenu GC et la couverture de lecture sur l'ensemble d'un génome. Les biais peuvent se produire lors de la préparation de librairies ou de la cartographie, dans les trousses Capture ou dans les différences entre les séquenceurs, et rendent difficile la définition de variants du nombre de copies (VNC). Le module de correction des biais GC de la plateforme DRAGEN tente de corriger ces biais.

Le module suit immédiatement l'étape de dénombrements de cibles et est exécuté sur le fichier .target.count. La correction de biais GC génère une version corrigée du fichier et ajoute une extension .target.counts.gc-corrected à son nom. Il est recommandé d'utiliser les versions pour lesquelles les biais GC ont été corrigés pour tout traitement en aval lorsque vous travaillez avec des données de séquençage de génome entier. Pour le séquençage d'exome entier, s'il y a suffisamment de régions cibles, alors les dénombrements pour lesquels les biais GC ont été corrigés peuvent aussi être utilisés.

Les trousses Capture classiques comprennent plus de 200 000 cibles dans les régions d'intérêt. Si votre fichier BED contient moins de 200 000 cibles, ou si les régions cibles se situent dans une région précise du génome (et que les statistiques de biais GC peuvent être décalées), alors la correction de biais GC devrait être désactivée.

Les options suivantes contrôlent le module de correction de biais GC.

- *--cnv-enable-gcbias-correction* – Permet d'activer ou de désactiver la correction de biais GC lors de la génération des dénombrements de cibles. La valeur par défaut est « true » (activée).

- *--cnv-enable-gcbias-smoothing* – Permet d'activer ou de désactiver la correction de biais GC simplifiée au sein de compartiments GC ayant un noyau exponentiel. La valeur par défaut est « true » (activée).
- *--cnv-num-gc-bins* – Spécifie le nombre de compartiments pour la correction de biais GC. Chaque compartiment représente un pourcentage de contenu GC. Les valeurs acceptées sont « 10 », « 20 », « 25 », « 50 » et « 100 ». La valeur par défaut est « 25 ».

Normalisation

Le pipeline de variants du nombre de copies (VNC) de la plateforme DRAGEN prend en charge deux algorithmes de normalisation :

- Normalisation automatique – Un algorithme qui utilise des statistiques de l'échantillon en cours d'analyse pour établir les niveaux de ploïdie de référence.
- Panel de référence – Un algorithme de normalisation basé sur la référence qui utilise des échantillons de référence correspondants supplémentaires pour établir un niveau de référence à partir duquel la définition des événements de VNC pourra être effectuée. Dans ce cas, « échantillons de référence correspondants » sous-entend que le même flux de travail de préparation de librairies et de séquençage que l'échantillon à analyser a été utilisé.

L'algorithme à utiliser dépend des données disponibles et de l'application. Suivez les recommandations suivantes pour sélectionner le mode de normalisation.

Normalisation automatique

- Séquençage de génome entier
- Analyse uniéchantillon
- Les échantillons correspondants supplémentaires ne sont pas faciles à obtenir
- Flux de travail plus simple avec une seule invocation

Panel de référence

- Séquençage de génome entier
- Séquençage d'exome entier
- Panels ciblés
- Les échantillons correspondants supplémentaires sont faciles à obtenir
- Analyse d'échantillons de tumeur et de référence correspondants

Normalisation automatique

Le pipeline de VNC de la plateforme DRAGEN offre le mode de normalisation automatique qui ne nécessite pas d'échantillon de référence ou de panel de référence. Ce mode peut être activé en mettant à l'option *--cnv-enable-self-normalization* la valeur « true » (activée). Ce mode de fonctionnement élimine deux étapes et permet donc d'économiser du temps. Il utilise les statistiques de l'échantillon à analyser pour établir le niveau de référence à partir duquel est effectuée une définition.

Comme la normalisation automatique se sert des statistiques de l'échantillon à analyser, ce mode n'est pas recommandé pour le séquençage du génome entier ou l'analyse de séquençage ciblé, car les données pourraient être insuffisantes.

Le mode de normalisation automatique est recommandé pour le séquençage de génome entier uniéchantillon. Le pipeline se poursuit avec les étapes de segmentation et de définition pour produire les événements définis finaux.

```
dragen \
   -r <HASHTABLE> \
   --bam-input <BAM> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
   --enable-map-align false \
   --enable-cnv true \
   --cnv-enable-self-normalization true
```

Si vous analysez un échantillon d'un fichier FASTQ, le mode de fonctionnement par défaut est la normalisation automatique.

Quand le mode de normalisation automatique est utilisé, la valeur de l'option *--cnv-interval-width* qui a servie lors de l'étape des dénombrements de cibles correspond à la largeur de l'intervalle en fonction du nombre de positions de k-mers uniques. Vous n'avez habituellement pas besoin de modifier cette option.

Panel de référence

L'approche avec panel de référence utilise un ensemble d'échantillons de référence correspondants pour déterminer le niveau de référence à partir duquel seront définis les événements de variants du nombre de copies. Ces échantillons de référence correspondants doivent être dérivés grâce aux mêmes flux de travail de préparation de librairies et de séquençage que ceux de l'échantillon à analyser. Ceci permet à l'algorithme de corriger les biais systématiques qui ne sont pas propres à l'échantillon.

Dans ce mode de fonctionnement, le pipeline de variants du nombre de copies de la plateforme DRAGEN est divisé en deux étapes distinctes. L'étape des dénombrements de cibles est effectuée sur chaque échantillon, celui à analyser et ceux de référence, pour classer les alignements. L'étape de normalisation et de détection des définitions est ensuite effectuée en comparant l'échantillon à analyser au panel de référence afin de définir les événements.

Étape des dénombrements de cibles

L'étape des dénombrements de cibles devrait être effectuée pour tous vos échantillons, que se soient des échantillons de référence ou des échantillons à analyser. L'échantillon à analyser et tous les échantillons à utiliser comme panel de référence doivent avoir des intervalles identiques et devraient donc être générés à partir de paramètres identiques. L'étape des dénombrements de cibles effectue également la correction du biais de GC, qui est activée par défaut.

Les exemples ci-dessous concernent le séquençage du génome entier. Pour le traitement des exomes, consultez la section *Exome entier* à la page 63.

L'exemple suivant montre une commande de traitement d'un fichier BAM.

```
dragen \
   -r <HASHTABLE> \
   --bam-input <BAM> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE. \
   --enable-map-align false \
   --enable-cnv true
```

Ci-dessous se trouve un exemple de commande pour le traitement d'un fichier CRAM.

```
dragen \
   -r <HASHTABLE> \
   --cram-input <CRAM> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
   --enable-map-align false \
   --enable-cnv true
```

Étape de normalisation et de détection des définitions

La prochaine étape du pipeline de variants du nombre de copies quand vous utilisez un panel de référence est d'effectuer la normalisation et les définitions. Cette étape comprend la sélection d'un panel de référence, une liste de fichiers de dénombrements de cibles servant à la normalisation de la médiane basée sur la référence.

Vous pouvez effectuer l'analyse dans d'autres combinaisons de flux de travail, mais n'oubliez pas que les événements de variants du nombre de copies sont définis pour les échantillons de référence utilisés. Idéalement, les échantillons du panel de référence suivent les mêmes flux de travail de préparation de librairies et de séquençage que ceux de l'échantillon en cours d'analyse. Pour la définition sur les chromosomes sexuels, il est recommandé d'utiliser des références du même sexe dans le panel. Comme la normalisation est effectuée par cible en effectuant la comparaison avec la médiane du panel de référence, avoir des références du même sexe est essentiel pour détecter les événements de nombre de copies sur les chromosomes sexuels.

Pour générer un panel de référence, créez un fichier texte brut dans lequel chaque ligne contient un chemin d'accès pointant à un fichier target.counts généré lors de l'étape des dénombrements de cibles. Les chemins d'accès relatifs sont pris en charge s'ils sont relatifs au répertoire de travail actuel. Les chemins d'accès absolus sont recommandés si le flux de travail sera utilisé ultérieurement ou partagé avec d'autres utilisateurs.

L'exemple suivant montre un fichier de panel de référence dans lequel est utilisé un sous-ensemble des fichiers avec GC corrigée provenant de l'étape des dénombrements de cibles.

```
/data/output_trio1/sample1.target.counts.gc-corrected
/data/output_trio1/sample2.target.counts.gc-corrected
/data/output_trio2/sample4.target.counts.gc-corrected
/data/output_trio2/sample5.target.counts.gc-corrected
/data/output_trio3/sample7.target.counts.gc-corrected
/data/output_trio3/sample8.target.counts.gc-corrected
```

Les fichiers à utiliser dans le panel de référence peuvent également être spécifiés grâce à l'option *--cnv-normals-file*. Cette option ne peut avoir qu'un seul nom de fichier en paramètre, mais peut être spécifiée plusieurs fois.

Après avoir créer un fichier de panel de référence, vous pouvez exécuter la fonction de définition en spécifiant votre échantillon à analyser grâce à l'option *--cnv-input* et le fichier de panel de référence grâce à l'option *--cnv-normals-list*. Comme nous recommandons d'utiliser les dénombrements avec correction du biais de GC de l'étape précédente, il n'est pas nécessaire d'effectuer de nouveau la correction du biais de GC. La correction du biais de GC peut être désactivée en mettant à l'option *--cnv-enable-gcbias-correction* la valeur « false » (désactivée). Par exemple :

```
dragen \
   -r <HASHTABLE> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
```

```
--enable-map-align false \
--enable-cnv true \
--cnv-input <CASE_COUNTS> \
--cnv-normals-list <NORMALS> \
--cnv-enable-gcbias-correction false
```

Cette commande normalise l'échantillon à analyser en effectuant la comparaison avec le panel de référence, puis passe aux étapes de segmentation et de définition en aval.

Options relatives à la normalisation

Ces options spécifient le préconditionnement du panel de référence et la normalisation de l'échantillon à analyser.

- *--cnv-enable-self-normalization* – Active ou désactive le mode de normalisation automatique. Celle-ci ne nécessite pas de panel de référence.
- *--cnv-extreme-percentile* – Spécifie la valeur seuil de percentile médian utilisée pour exclure des échantillons. La valeur par défaut est « 2,5 ».
- *--cnv-input* – Spécifie un fichier de dénombrements de cibles pour l'échantillon à analyser quand un panel de référence est utilisé.
- *--cnv-matched-normal* – Spécifie le fichier de dénombrements de cibles de l'échantillon de référence correspondant.
- *--cnv-normals-file* – Spécifie le fichier target.counts à utiliser dans le panel de référence. Cette option peut être spécifiée plusieurs fois, une par fichier.
- *--cnv-normals-list* – Spécifie le fichier texte contenant les chemins d'accès de la liste des fichiers de dénombrements de cibles de référence à utiliser en tant que panel de référence. Les chemins d'accès absolus sont recommandés si le flux de travail sera utilisé ultérieurement ou partagé avec d'autres utilisateurs. Les chemins d'accès relatifs sont pris en charge s'ils sont relatifs au répertoire de travail actuel.
- *--cnv-max-percent-zero-samples* – Spécifie un seuil pour le filtrage de cibles avec trop d'échantillons sans couverture. La valeur par défaut est « 5 % ».
- *--cnv-max-percent-zero-targets* – Spécifie un seuil pour le filtrage d'échantillons avec trop de cibles sans couverture. La valeur par défaut est « 2,5 % ».
- *--cnv-target-factor-threshold* – Spécifie un pourcentage de seuil de facteur cible médian pour exclure les cibles utilisables. La valeur par défaut est « 1 % » pour le traitement du génome entier et « 10 % » pour le traitement du séquençage ciblé.
- *--cnv-truncate-threshold* – Spécifie un pourcentage de seuil pour la troncature des valeurs aberrantes extrêmes. La valeur par défaut est « 0,1 % ».

Segmentation

Une fois qu'un échantillon à analyser a été normalisé, il passe à l'étape de la segmentation.
Plusieurs algorithmes de segmentation sont intégrés à la plateforme DRAGEN, notamment :

- CBS (segmentation binaire circulaire)
- SLM (modèles multiéchelles)
- FPOP (partitionnement optimal et élagage fonctionnel)

L'algorithme SLM a deux variants : SLM et HSLM. L'algorithme HSLM (modèles multiéchelles hétérogènes) est utilisé aux fins d'analyse de l'exome et traite des trousses de capture de cibles dont l'espacement est inégal.

Les algorithmes de segmentation par défaut sont l'algorithme SLM pour le traitement d'un génome entier et l'algorithme CBS pour un exome entier.

- *--cnv-segmentation-mode* – Spécifie l'algorithme de segmentation à utiliser. Les valeurs possibles sont « cbs », « slm », « hslm », ou « fpop ». La valeur par défaut est « slm » ou « cbs », selon si les intervalles sont des intervalles de génomes entiers ou des intervalles de séquençage ciblé.
- *--cnv-merge-distance* – Spécifie le nombre minimal de paires de bases entre deux segments afin de permettre leur fusion. La valeur par défaut est « 0 », c'est-à-dire que les segments doivent être directement adjacents.
- *--cnv-merge-threshold* – Spécifie la différence maximale entre les valeurs moyennes des segments à laquelle deux segments adjacents peuvent être fusionnés. La valeur moyenne des segments est représentée par un rapport de copies linéaire. La valeur par défaut est « 0,2 ». La valeur 0 désactive la fusion.

Ensembles R externes

Pour utiliser les algorithmes de segmentation SLM ou FPOP, vous devez installer des ensembles R externes. Les ensembles R sont exécutés à l'extérieur du logiciel hôte de la plateforme DRAGEN. Un simple script d'installation est intégré à la plateforme : /opt/edico/R/install_R_packages.R. Ces ensembles R sont des méthodes de segmentation supplémentaires qui peuvent être utilisées, mais qui ne sont pas officiellement gérées par le logiciel hôte de la plateforme DRAGEN. Si votre système ne permet pas l'installation d'ensembles R ou de librairies tierces, alors vous pouvez utiliser la segmentation binaire circulaire par défaut.

Segmentation binaire circulaire

La segmentation binaire circulaire est mise en œuvre directement dans la plateforme DRAGEN et est basée sur l'article *A faster circular binary segmentation for the analysis of array CGH data* (en anglais). Elle comprend des améliorations permettant d'augmenter la sensibilité pour les données de séquençage nouvelle génération (SNG).

Les options suivantes contrôlent la segmentation binaire circulaire.

- *--c-alpha* – Spécifie le niveau d'importance à partir duquel le test accepte les modifications. La valeur par défaut est « 0,01 ».
- *--cnv-cbs-eta* – Spécifie le taux d'erreur de type I des limites séquentielles pour un arrêt précoce lorsque la méthode de permutation est utilisée. La valeur par défaut est « 0,05 ».
- *--cnv-cbs-kmax* – Spécifie la largeur maximale du plus petit segment destiné à la permutation. La valeur par défaut est « 25 ».
- *--cnv-cbs-min-width* – Spécifie le nombre minimal de marqueurs pour un segment modifié. La valeur par défaut est « 2 ».
- *--cnv-cbs-nmin* – Spécifie la longueur minimale de données pour une approximation statistique maximale. La valeur par défaut est « 200 ».
- *--cnv-cbs-nperm* – Spécifie le nombre de permutations utilisées dans le calcul de la valeur p. La valeur par défaut est « 10 000 ».
- *--cnv-cbs-trim* – Spécifie la proportion des données à retrancher aux fins de calculs de la variance. La valeur par défaut est « 0,025 ».

Segmentation SLM (modèles multiéchelles)

Le mode de segmentation SLM suit l'intégration R présentée dans l'article SLMSuite: a suite of algorithms for segmenting genomic profiles (en anglais).

- *--cnv-slm-eta* – La probabilité de référence que la valeur du processus moyen change. La valeur par défaut est « 1e-5 ».
- *--cnv-slm-fw* – Le nombre minimal de points de données pour produire un variant du nombre de copies. La valeur par défaut est « 0 », ce qui signifie que les segments comportant une conception de sonde pourraient dans les faits être produits.
- *--cnv-slm-omega* – Le paramètre d'échelle modulant la pondération relative entre la variance expérimentale/biologique. La valeur par défaut est « 0,3 ».
- *--cnv-slm-stepeta* – Le paramètre de normalisation de la distance. La valeur par défaut est « 10 000 ». Cette option est valide seulement pour le mode HSLM.

Segmentation FPOP (partitionnement optimal et élagage fonctionnel)

Le mode de segmentation FPOP suit l'intégration R présentée dans l'article On Optimal Multiple Changepoint Algorithms for Large Data (en anglais).

- *--cnv-fpop-penalty* – Option de pénalité pour la détection des modifications. La valeur par défaut est « 0,03 ».

Notation de la qualité

Les scores de qualité sont calculés à l'aide d'un modèle probabiliste qui utilise un mélange de distributions de probabilité à queue lourde (une par nombre entier de copies) et une pondération relative à la longueur des événements. La variance du bruit est estimée. Le fichier VCF de sortie contient un indicateur de l'échelle Phred qui mesure la confiance dans les événements d'amplification définis (nombre de copies > 2 pour les locus diploïdes) ou de nombre de copies neutre (nombre de copies = 2 pour les locus diploïdes).

L'algorithme de notation calcule aussi les scores de qualité du nombre exact de copies des fichiers d'entrée du pipeline de détection des VNC de novo.

Fichiers de sortie

Le logiciel hôte de la plateforme DRAGEN génère plusieurs fichiers intermédiaires. Le fichier de définitions final contenant les événements d'amplification et de délétion est le fichier *.seg.called.merged.

En plus du fichier de segments, la plateforme DRAGEN produit les définitions dans le format VCF standard. Par défaut, le fichier VCF comprend seulement les événements de gains et de pertes du nombre de copies. Pour les segments au nombre de copies neutre, consultez le fichier *.seg.called.merged. Pour que les définitions au nombre de copies neutre (REF) soient comprises dans le fichier VCF de sortie, mettez à l'option *--cnv-enable-ref-calls* la valeur « true » (activée).

Pour en savoir plus au sujet du fichier .seg.called.merged et pour savoir comment utiliser les fichiers de sortie aux fins de débogage et d'analyse, consultez la section *Analyse du flux relatif au signal* à la page 58.

Fichier VCF de variants du nombre de copies

Le fichier VCF de variants du nombre de copies (VNC) suit le format VCF standard. À cause de la nature de la représentation des événements relatifs aux VNC par rapport à celle des variants structurels, les champs ne s'appliquent pas tous. En général, si des renseignements sont disponibles au sujet d'un événement, alors ils sont annotés. Certains champs du fichier VCF de VNC sont propres aux VNC.

L'exemple suivant montre les lignes d'en-tête qui sont propres aux VNC :

```
##fileformat=VCFv4.1
##ALT=<ID=CNV,Description="Copy number variant region">
##ALT=<ID=DEL,Description="Deletion relative to the reference">
##ALT=<ID=DUP,Description="Region of elevated copy number relative to the
reference">
##contig=<ID=1,length=249250621>
##contig=<ID=2,length=243199373>
##contig=<ID=3,length=198022430>
##contig=<ID=4,length=191154276>
##contig=<ID=5,length=180915260>
…
##reference=file:///reference_genomes/Hsapiens/hs37d5/DRAGEN
##INFO=<ID=REFLEN,Number=1,Type=Integer,Description="Number of REF positions
included in this record">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant
described in this record">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around
POS">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around
END">
##FILTER=<ID=cnvQual,Description="CNV with quality below 10">
##FILTER=<ID=cnvCopyRatio,Description="CNV with copy ratio within +/- 0.2 of
1.0">
##FORMAT=<ID=SM,Number=1,Type=Float,Description="Linear copy ratio of the segment
mean">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Estimated copy number">
##FORMAT=<ID=BC,Number=1,Type=Integer,Description="Number of bins in the region">
##FORMAT=<ID=PE,Number=2,Type=Integer,Description="Number of improperly paired
end reads at start and stop breakpoints">
```

La colonne ID est utilisée pour représenter l'événement.

La colonne REF contient un « N » pour tous les événements relatifs aux VNC.

La colonne ALT indique le type d'événement relatif aux VNC. Puisque le fichier VCF ne contient que des événements relatifs aux VNC, les seules entrées utilisées sont DEL ou DUP.

La colonne QUAL contient un score de qualité estimé pour l'événement relatif aux VNC, qui est ensuite utilisé comme filtre.

La colonne FILTER contient la mention « PASS » (réussite) si l'événement relatif aux VNC passe tous les filtres. Autrement, cette colonne contient le nom du filtre que l'événement ne passe pas.

La colonne INFO contient des renseignements représentant l'événement, en grande partie identiques à ceux inscrits dans la colonne ID. L'entrée REFLEN indique la longueur de l'événement. L'entrée SVTYPE contient toujours « CNV ». L'entrée END indique la position finale de l'événement. Les entrées CIPOS et CIEND ne sont pas actuellement utilisées.

Les champs FORMAT sont décrits dans l'en-tête.

- SM – Le rapport de copies linéaire de la moyenne du segment.

- CN – L'estimation du nombre de copies.
- BC – Le nombre de compartiments dans la région.
- PE – Le nombre de lectures mal appariées aux points de rupture de départ et de fin.
- GT – Le génotype. Puisque la définition du nombre de copies germinales détermine le nombre de copies global plutôt que le nombre de copies sur chaque haplotype, le champ indiquant le type de génotype contient des valeurs manquantes pour les régions diploïdes lorsque la valeur de « CN » est supérieure ou égale à « 2 ». Les exemples suivants montrent le champ « GT » pour différentes entrées de fichier VCF :

Diploïde ou haploïde?	ALT	FORMAT:CN	FORMAT:GT
Diploïde	.	2	./.
Diploïde	<DUP>	> 2	./1
Diploïde	<DEL>	1	0/1
Diploïde	<DEL>	0	1/1
Haploïde	.	1	0
Haploïde	<DUP>	> 1	1
Haploïde	<DEL>	0	1

Visualisation et fichiers bigWig

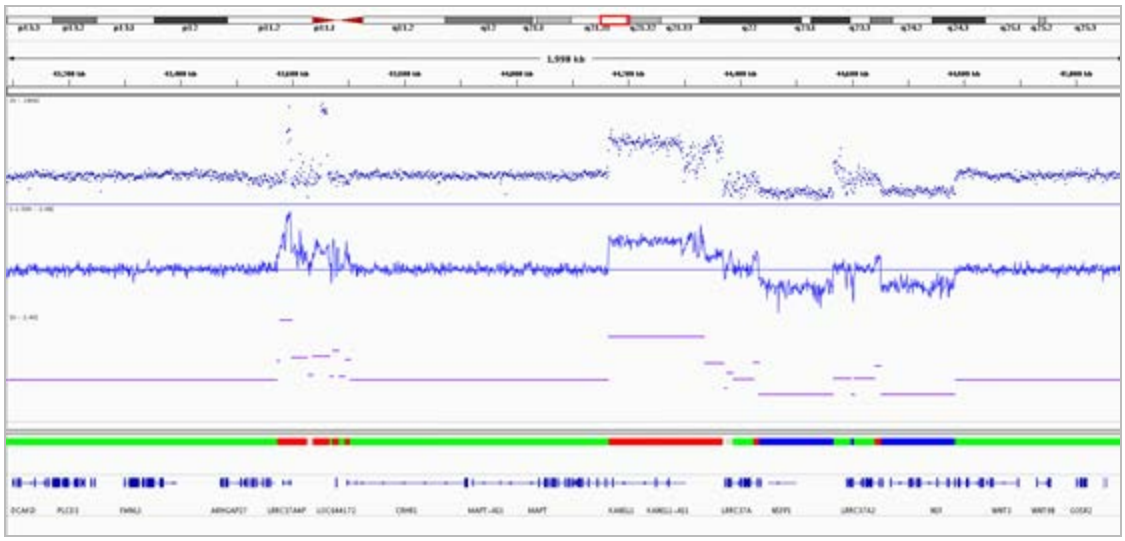
Pour effectuer une analyse sur un ensemble représentatif de la réalité connu, vous pouvez utiliser les fichiers de sortie intermédiaires provenant des étapes du pipeline. Ces fichiers peuvent être analysés pour peaufiner les options.

Tous les fichiers ont une structure similaire à celle d'un fichier BED, avec une ligne d'en-tête facultative.

- **.target.counts* – Contient le nombre de lectures par intervalle de cibles. Il s'agit du signal brut tel qu'il est extrait des alignements du fichier BAM ou CRAM. Le format est identique pour l'échantillon à analyser et n'importe quel échantillon d'un panel de référence. Une représentation bigWig est aussi effectuée dans un fichier target.counts.diploid, qui est normalisé avec le niveau de ploïdie 2 de référence plutôt que les dénombrements bruts.
- **.tn.tsv* – Le signal moyen normalisé de l'échantillon à analyser, par intervalle de cibles. Ce fichier contient le signal normalisé consigné. Un écart important du signal de « 0,0 » indique un potentiel événement relatif aux VNC.
- **.seg.called.merged* – Contient les segments produits par l'algorithme de segmentation.
- **.cnv.vcf* – Le fichier VCF de sortie indiquant des événements relatifs aux VNC.

Pour générer des fichiers supplémentaires équivalents aux formats bigWig et GFF, mettez à l'option *--enable-cnv-tracks* la valeur « true » (activée). Ces fichiers peuvent être chargés dans IGV avec d'autres pistes disponibles, par exemple les gènes RefSeq. L'utilisation de ces pistes, en plus des pistes publiques, permet une interprétation plus facile des définitions. La figure ci-dessous en montre un exemple :

Figure 9 Exemple d'IGV



Options de sortie et de filtres

Les options de sortie et de filtres contrôlent les fichiers de sortie des VNC.

- *--cnv-blacklist-bed* – Spécifie un fichier BED indiquant des intervalles à exclure du fichier de VNC de sortie final. Si une définition chevauche au moins 50 % de n'importe quel intervalle du fichier BED de régions à ignorer, elle est supprimée.
- *--cnv-enable-plots* – Génère des tracés dans le cadre du pipeline de VNC. La valeur par défaut est « false » (désactivée).
- Si vous effectuez une analyse de VNC du séquençage du génome entier avec des intervalles à haute résolution (moins de 1 000 paires de bases), alors la génération du tracé peut prendre plus de temps. Illumina vous recommande d'utiliser la valeur par défaut (option désactivée).
- *--cnv-enable-ref-calls* – Produit des définitions au nombre de copies neutre (REF) dans le fichier VCF de sortie. La valeur par défaut est « false » (désactivée).
- *--cnv-enable-tracks* – Génère des fichiers de suivi pouvant être importés dans l'outil Integrative Genome Viewer (IGV). Lorsque l'option est activée, un fichier *.gff est généré pour les définitions de variants de sortie, ainsi que des fichiers *.bw pour le signal moyen normalisé. La valeur par défaut est « true » (activée).
- *--cnv-filter-bin-support-ratio* – Filtre un événement candidat si l'étendue des compartiments connexes est inférieure au rapport spécifié conformément à la longueur globale de l'événement. Le rapport par défaut est « 0,2 » (20 %). Par exemple, si un événement est défini et a une longueur de 100 000 paires de bases, mais que les compartiments de l'intervalle de cibles qui soutiennent la définition ne couvrent qu'un total de 15 000 paires de bases (15 000/100 000 = 0,15), alors cet événement sera filtré.
- *--cnv-filter-copy-ratio* – Spécifie le rapport minimal du nombre de copies, centré autour de « 1,0 », à partir duquel un événement est marqué avec « PASS » (réussite) dans le fichier VCF de sortie. La valeur par défaut est « 0,2 », pour des définitions dont le rapport du nombre de copies est inférieur à « 0,8 » ou supérieur à « 1,2 ».
- *--cnv-filter-length* – Spécifie la longueur maximale d'un événement (dans les bases) pour laquelle un événement d'un rapport est marqué « PASS » (réussite) dans le fichier VCF de sortie. La valeur par défaut est « 10 000 ».

- *--cnv-filter-qual* – Spécifie la valeur QUAL pour laquelle un événement d'un rapport est marqué « PASS » (réussite) dans le fichier VCF de sortie. La valeur par défaut est « 10 ».
- *--cnv-min-qual* – Spécifie la valeur QUAL minimale à signaler. La valeur par défaut est « 3 ».
- *--cnv-max-qual* – Spécifie la valeur QUAL maximale à signaler. La valeur par défaut est « 200 ».
- *--cnv-ploidy* – Spécifie la valeur de ploïdie de référence. Cette option est utilisée pour l'estimation du nombre de copies produite dans le fichier VCF de sortie. La valeur par défaut est « 2 ».
- *--cnv-qual-length-scale* – Spécifie le facteur de pondération de biais visant à ajuster les estimations QUAL pour les segments plus longs. Il s'agit d'une option avancée qui ne devrait pas être modifiée. La valeur par défaut est « 0,9303 » (2-0,1).
- *--cnv-qual-noise-scale* – Spécifie le facteur de pondération de biais pour ajuster les estimations QUAL à partir de la variance de l'échantillon. Il s'agit d'une option avancée qui ne devrait pas être modifiée. La valeur par défaut est « 1,0 ».

Définition simultanée de variants du nombre de copies et de petits variants

La plateforme DRAGEN peut effectuer la cartographie et l'alignement d'échantillons FASTQ et transmettre les données aux fonctions de définition en aval. Un échantillon unique peut être analysé pour le variant du nombre de copies et pour la définition de petits variants si l'échantillon d'entrée est en format FASTQ. Cela déclenche la normalisation automatique par défaut.

L'échantillon FASTQ est analysé selon un flux de travail normal de cartographie et d'alignement par la plateforme DRAGEN, puis il est possible d'ajouter des arguments supplémentaires pour permettre la définition de VNC, la définition de variants, ou les deux. Les options qui s'appliquent à la définition de VNC dans les flux de travaux autonomes s'appliquent également ici.

Les exemples suivants présentent différentes commandes.

Cartographie et alignement d'un échantillon FASTQ avec définition de VNC

```
dragen \
   -r <HASHTABLE> \
   -1 <FASTQ1> \
   -2 <FASTQ2> \
   --RGSM <SAMPLE> \
   --RGID <RGID> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
   --enable-map-align true \
   --enable-cnv true \
   --cnv-enable-self-normalization true
```

Cartographie et alignement d'un échantillon FASTQ avec définition de variants

```
dragen \
   -r <HASHTABLE> \
   -1 <FASTQ1> \
   -2 <FASTQ2> \
   --RGSM <SAMPLE> \
   --RGID <RGID> \
   --output-directory <OUTPUT> \
```

```
    --output-file-prefix <SAMPLE> \
    --enable-map-align true \
    --enable-variant-caller true
```

Cartographie et alignement d'un échantillon FASTQ avec définition de VNC et définition de variants

```
dragen \
   -r <HASHTABLE> \
   -1 <FASTQ1> \
   -2 <FASTQ2> \
   --RGSM <SAMPLE> \
   --RGID <RGID> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
   --enable-map-align true \
   --enable-cnv true \
   --cnv-enable-self-normalization true \
   --enable-variant-caller true
```

Fichier BAM d'entrée pour définition de VNC et définition de variants

```
dragen \
   -r <HASHTABLE> \
   --bam-input <BAM> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
   --enable-map-align false \
   --enable-cnv true \
   --cnv-enable-self-normalization true \
   --enable-variant-caller true
```

Corrélation des échantillons et génotypage du sexe

À l'étape de dénombrements de cibles ou de la normalisation, le pipeline de VNC de la plateforme DRAGEN fournit également les renseignements suivants au sujet des échantillons analysés :

- Un indicateur de corrélation du profil de dénombrement de lectures entre l'échantillon analysé et un échantillon d'un panel de référence. Un indicateur de corrélation supérieur à « 0,90 » est recommandé pour une analyse de confiance, mais le logiciel n'applique aucune restriction.
- Une prédiction du sexe de chaque échantillon analysé. Le sexe est prédit à partir des renseignements sur le dénombrement de lectures dans les chromosomes sexuels et les chromosomes autosomiques. La valeur médiane pour ces nombres s'affiche à l'écran pour les chromosomes autosomiques, le chromosome X et le chromosome Y.

Les résultats s'affichent à l'écran lors de l'exécution du pipeline. Par exemple :

```
==========================================
Correlation Table
==========================================
Correlation of case sample PlatinumGenomes_50X_NA12877 against
PlatinumGenomes_50X_NA12878: 0.984092
```

```
Sex Genotyper
==========================================
Predicted sex of samples
PlatinumGenomes_50X_NA12877: MALE XY 0.99737
PlatinumGenomes_50X_NA12878: FEMALE XX 0.968929
```

Pour effectuer l'analyse des chromosomes sexuels à l'aide d'un panel de référence, nous vous recommandons d'utiliser des échantillons dont le sexe est le même que celui du panel de référence.

Vous pouvez remplacer le sexe de l'échantillon en utilisant l'option *–sample-sex*.

Définition de variants du nombre de copies multiéchantillons

Il est possible d'effectuer la définition de VNC multiéchantillons en utilisant les fichiers de dénombrements moyens normalisés (*.tn.tsv) spécifiés à l'aide de l'option *--cnv-input* (un par échantillon). L'analyse multiéchantillons des VNC tire parti de l'utilisation de la segmentation combinée pour améliorer la sensibilité de la détection des segments à nombre de copies variable. Pour chaque segment à nombre de copies variable identifié, le génotype du nombre de copies de chaque échantillon est produit dans une entrée de fichier VCF afin de faciliter l'annotation et l'interprétation.

L'exemple suivant montre une ligne de commande pour exécuter une analyse de trio :

```
dragen \
-r <HASHTABLE> \
--output-directory <OUTPUT> \
--output-file-prefix <SAMPLE> \
--enable-cnv true \
--cnv-input <FATHER_TN_TSV> \
--cnv-input <MOTHER_TN_TSV> \
--cnv-input <PROBAND_TN_TSV> \
--pedigree-file <PEDIGREE_FILE>
```

Options relatives à la définition de VNC de novo

Les options suivantes sont utilisées pour la définition de VNC de novo :

--cnv-input – Pour la définition de VNC, cette option spécifie les fichiers de signal moyen normalisé (*.tn.tsv) des analyses uniéchantillon. Cette option peut être spécifiée plusieurs fois, une par échantillon d'entrée.

--cnv-filter-de-novo-qual – Le seuil de l'échelle Phred à partir duquel l'événement putatif de l'échantillon du proposant est marqué comme de novo. La valeur par défaut est « 0,10 ».

--pedigree-file – Le fichier de généalogie qui spécifie la relation entre les échantillons d'entrée.

Segmentation combinée

La définition de variants du nombre de copies est d'abord effectuée indépendamment sur chaque échantillon. La segmentation combinée utilise ensuite les segments à nombre de copies variables de chaque analyse uniéchantillon pour dériver un ensemble combiné de segments à nombre de copies variables. Cet ensemble de segments combinés est défini simplement en prenant l'union de tous les points de rupture des segments à nombre de copies variables de tous les échantillons. Ceci entraîne le fractionnement de tous les segments se chevauchant partiellement dans les différents échantillons. Par exemple :

Après la segmentation combinée, la définition du nombre de copies est de nouveau effectuée indépendamment sur chaque échantillon en utilisant les segments combinés. Les segments peuvent être fusionnés comme pour l'analyse uniéchantillon, mais chaque segment combiné est ajouté en tant qu'une seule entrée dans le fichier VCF multiéchantillons. Le score de qualité (champ QS dans le fichier VCF) du segment fusionné de l'échantillon, le cas échéant, sert à filtrer la définition. Les définitions de l'échantillon sont filtrées à l'aide du champ FT de l'échantillon dans le fichier VCF multiéchantillons. La colonne QUAL du fichier VCF n'a jamais de valeur (c.-à-d. que la valeur est « . »). La colonne FILTER du fichier VCF multiéchantillons reçoit la valeur « SampleFT » si aucun des champs FT de l'échantillon ne contient la valeur « PASS » (réussite). Elle reçoit la valeur « PASS » (réussite) si au moins un des champs FT de l'échantillon contient la valeur « PASS » (réussite).

Étape de la définition de novo

Un événement de novo est défini par l'existence d'un génotype à un locus particulier dans le génome du proposant qui ne résulte pas d'une hérédité mendélienne standard des parents. L'étape de définition de novo permet d'identifier les événements de novo putatifs, chez le proposant, de chaque trio d'une analyse multiéchantillons. Dans certains cas, ces événements de novo putatifs peuvent être réels, mais ils peuvent également être produits par des artéfacts de séquençage ou d'analyse. Par conséquent, un score de qualité de novo est attribué à chaque événement de novo putatif et est utilisé pour filtrer les événements de novo de mauvaise qualité. Les trios sont définis en spécifiant un fichier .ped avec l'option *--pedigree-file*. Plusieurs trios peuvent être spécifiés (p. ex., une analyse de quadrants), mais seuls les trios valides seront traités.

Pour chaque segment combiné d'un trio, la définition de novo détermine s'il y a un conflit d'hérédité mendélienne pour les génotypes de nombre de copies définis. La fonction de définition de variants du nombre de copies ne détecte pas le nombre de copies pour chaque allèle d'un segment diploïde donné, ce qui signifie que des suppositions sont effectuées à propos de la composition allélique possible des génotypes parentaux.

La supposition est qu'il n'y a aucun allèle présentant zéro copie dans les régions diploïdes du génome d'un parent (selon le sexe) quand le nombre de copies attribuées est égal ou supérieur à 2. Ceci engendre des simplifications, comme suit :

Génotype de nombre de copies du parent	Allèles de nombre de copies possibles	Allèles de nombre de copies possibles supposés
2	0/2, 1/1	11
3	0/3, 1/2	1/2
4	0/4, 1/3, 2/2	1/3, 2/2
N	x/(N-x) pour x <= N/2	x/(N-x) pour 1 <=x <= N/2

Ci-dessous se trouvent des exemples cohérents et incohérents de génotypes de nombre de copies des régions diploïdes utilisant ces suppositions :

Nombre de copies de la mère	Nombre de copies du père	Nombre de copies du proposant	Cohérence mendélienne?
2	2	2	Oui
2	2	1	Non
3	2	4	Non
3	2	2	Oui
2	0	2	Non

S'il y a un conflit d'hérédité mendélienne dans un segment combiné, un score de qualité de novo de l'échelle Phred (champ DQ dans le fichier VCF) est calculé en utilisant les probabilités pour chaque statut de nombre de copies (consultez la section Notation de qualité) de chacun des échantillons du trio et la probabilité à priori pour le génotype du trio :

```
DQ = -10log(1-Sum over conflicting genotypes(p(CNm|data)*p(CNf|data)*p
   (CNp|data)*p(CNm,CNf,CNp))/Sum over all genotypes(p(CNm|data)*p
   (CNf|data)*p(CNp|data)*p(CNm,CNf,CNp)))
```

Où

- CNm = Nombre de copies de la mère
- CNf = Nombre de copies du père
- CNp = Nombre de copies du proposant
- p(CNm,CNf,CNp) = la probabilité à priori du génotype du trio

Le champ DN du fichier VCF sert à indiquer le statut de novo pour chaque segment. Les valeurs possibles sont :

- Inherited (Transmis) – Le typage génotypique du trio est cohérent avec l'hérédité mendélienne.
- LowDQ (Qualité de novo faible) – Le typage génotypique du trio n'est pas cohérent avec l'hérédité mendélienne et la qualité de novo est inférieure au seuil de qualité de novo (« 0,1 » par défaut).
- DeNovo (Qualité de novo) – Le typage génotypique du trio n'est pas cohérent avec l'hérédité mendélienne et la qualité de novo est supérieure ou égale au seuil de qualité de novo (« 0,1 » par défaut).

Fichier VCF de sortie de variants du nombre de copies multiéchantillons

Les enregistrements dans un fichier VCF de variants du nombre de copies multiéchantillons sont légèrement différents de ceux d'un fichier uniéchantillon. Les principales différences sont les suivantes :

- Les entrées par enregistrement sont réparties en segments parmi l'union des points de rupture des échantillons en entrée, ce qui signifie qu'il y a plus d'entrées dans l'ensemble du fichier VCF.
- La colonne QUAL n'est pas utilisée et sa valeur est « . ». La qualité par échantillon est reportée dans les colonnes SAMPLE et comprend l'étiquette QS.
- La colonne FILTER (filtre) a la valeur « PASS » (réussite) si la valeur d'au moins une des colonnes SAMPLE est « PASS » (réussite). Sinon, la valeur est « SampleFT ».
- Les annotations par échantillon sont reportées depuis leurs définitions d'origine. Les filtres uniéchantillon sont appliqués au niveau de l'échantillon et sont ajoutés dans l'annotation FT.

De plus, si une généalogie valide est utilisée, la définition de novo est effectuée, ce qui ajoute les deux annotations suivantes à l'échantillon du proposant.

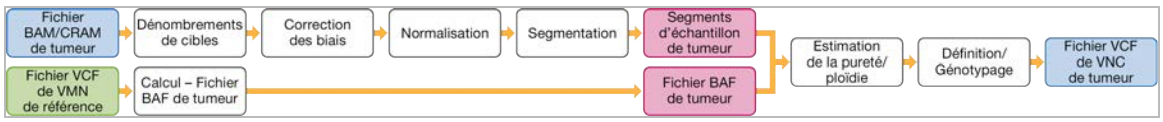
```
##FORMAT=<ID=DQ,Number=1,Type=Float,Description="De novo quality">
```

```
##FORMAT=<ID=DN,Number=1,Type=String,Description="Possible values are
   'Inherited', 'DeNovo' or 'LowDQ'. Threshold for a passing de novo call
   is DQ > 0.100000">
```

Même si le fichier VCF contient plusieurs entrées, en raison de l'étape de segmentation combinée, le nombre d'événements de novo peut être déterminé en extrayant les entrées qui ont une annotation DN et DQ. Ces enregistrements sont également extraits et convertis au format GFF3 si la définition de novo est utilisée.

Définition de variants somatiques du nombre de copies

Pour détecter les aberrations du nombre de copies, exécutez la fonction de définition de variants du nombre de copies sur une paire d'échantillons correspondants de génome entier de tumeur et de référence. Les composants de la fonction de définition de variants germinaux du nombre de copies sont réutilisés par l'algorithme somatique avec un composant de modélisation somatique qui estime la pureté tumorale et la ploïdie.



L'échantillon de référence correspondant est d'abord traité par la fonction de définition de petits variants germinaux pour déceler les sites de polymorphismes mononucléotidiques hétérozygotes. La fonction de définition de variants somatiques du nombre de copies se sert ensuite du fichier de sortie de la fonction de définition de petits variants de référence de l'échantillon pour mesurer le rapport d'allèle B dans l'échantillon de tumeur.

Le rapport d'allèle B permet la définition de variants du nombre de copies propre à l'allèle dans l'échantillon de tumeur. Le fichier de sortie est un fichier VCF. Les régions avec une perte d'hétérozygotie sont également détectées.

Options relatives à la définition de variants somatiques du nombre de copies

- *--tumor-bam-input*
- *--tumor-cram-input*
- *--cnv-normal-b-allele-vcf* ou *--cnv-population-b-allele-vcf*
- *--sample-sex*

Pour lancer la fonction de définition de variants somatiques du nombre de copies, les alignements en entrée doivent utiliser les options équivalentes pour l'échantillon de tumeur seulement comme *--tumor-bam-input* ou *--tumor-cram-input*. La fonction de définition de variants somatiques du nombre de copies ne prend pas en charge les fichiers FASTQ en entrée (*--tumor-fastq1*, *--tumor-fastq2* et *--tumor-fastq-list*). La fonction de définition de variants somatiques du nombre de copies ne prend pas non plus en charge l'exécution directement à partir de la fonction de cartographie et d'alignement.

Utilisez l'option *--cnv-normal-b-allele-vcf* pour spécifier un fichier VCF de variants mononucléotidiques de référence correspondant ou l'option *--cnv-population-b-allele-vcf* pour spécifier un catalogue de polymorphismes mononucléotidiques de population. Pour en savoir plus sur la façon de spécifier le locus de l'allèle B, consultez la section *Spécification du locus de l'allèle B* à la page 80.

Si le sexe associé à l'échantillon est connu, utilisez l'option *--sample-sex* pour spécifier le sexe. S'il n'est pas spécifié, la fonction de définition tentera de déduire le sexe associé à l'échantillon à partir des alignements.

L'exemple suivant montre une ligne de commande permettant d'exécuter la fonction de définition de variants somatiques du nombre de copies.

```
dragen \
   -r <HASHTABLE> \
   --output-directory <OUTPUT> \
   --output-file-prefix <SAMPLE> \
   --enable-map-align false \
   --enable-cnv true \
   --tumor-bam-input <TUMOR_BAM> \
   --cnv-normal-b-allele-vcf <SNV_NORMAL_VCF> \
   --sample-sex <SEX>
```

Dénombrements de cibles et d'allèles B

L'étape des dénombrements de cibles et la sortie sont les mêmes que pour la définition de variants germinaux du nombre de copies. Les intervalles cible sont produits avec les dénombrements de lectures dans un fichier *.target.counts. Le dénombrement d'allèles B est effectué en parallèle à l'étape de dénombrement de lectures et les valeurs sont produites dans un fichier *.baf.bedgraph.gz. Ce fichier peut être chargé dans IGV avec les autres fichiers BigWig générés par la plateforme DRAGEN aux fins de visualisation.

Spécification du locus de l'allèle B

Utilisez l'option *--cnv-normal-b-allele-vcf* pour spécifier un fichier VCF de variants de polymorphismes mononucléotidiques de référence correspondant. Idéalement, ce fichier VCF provient du traitement avec filtres de l'échantillon de référence correspondant par la fonction de définition de petits variants de la plateforme DRAGEN. Le nom de ce fichier a habituellement une extension *.hard-filtered.vcf.gz. Tous les enregistrements portant la marque « PASS » (réussite) qui ont été identifiés comme hétérozygotes dans l'échantillon de référence sont utilisés pour mesurer le dénombrement d'allèles B de l'échantillon de tumeur. Vous pouvez également utiliser le fichier gVCF équivalent (*.hard-filtered.gvcf.gz), mais la durée du traitement est beaucoup plus longue en raison du nombre important d'enregistrements, dont la plupart ne sont pas des sites hétérozygotes. Il est recommandé d'utiliser le fichier VCF.

Utilisez l'option *--cnv-population-b-allele-vcf* pour spécifier un fichier VCF de polymorphismes mononucléotidiques de population. Pour obtenir un fichier VCF de polymorphismes mononucléotidiques de population, traitez un catalogue approprié de variations de population, comme ceux provenant de dbSNP, du projet 1000G ou de toute autre initiative de recherche de grandes cohortes. Seuls les polymorphismes mononucléotidiques à fréquence élevée devraient être inclus. Par exemple, incluez les polymorphismes mononucléotidiques avec une fréquence d'allèle mineur de population égale ou supérieur à 10 % afin de limiter la durée de l'exécution et réduire les artéfacts. Spécifiez la fréquence de l'allèle ALT en ajoutant la fréquence à la valeur du champ AF = <alt frequency> de la section INFO de chaque enregistrement. Il pourrait y avoir des champs supplémentaires dans la section INFO, mais la plateforme DRAGEN analyse la chaîne pour n'utiliser que la valeur du champ AF. Les sites spécifiés par l'option *--cnv-population-b-allele-vcf* peuvent être hétérozygotes ou homozygotes dans le génome germinal duquel le génome tumoral est dérivé.

Ci-dessous se trouve un exemple d'enregistrement valide de polymorphisme mononucléotidique de population :

```
chr1 51479 . T A 1000 PASS AF=0.3253
```

La plateforme DRAGEN prend en considération les exigences suivantes lorsqu'elle analyse les enregistrements du fichier VCF d'allèle B :

- Sites de variants mononucléotidiques seulement.

- Les enregistrements doivent avoir la valeur « PASS » (réussite) dans le champ FILTER.
- S'il y a des enregistrements en double (avec les mêmes valeurs dans les champs CHROM et POS) dans le fichier VCF, alors seul le premier enregistrement est utilisé.

Modèle de variants somatiques du nombre de copies

Sélectionner un niveau de pureté de tumeur et de couverture diploïde (ploïdie) est un élément clé de la fonction de définition de variants somatiques du nombre de copies. Il tente de correspondre aux données observées, aux dénombrements de lectures et aux dénombrements d'allèle B dans tous les segments de l'échantillon de tumeur. Un score de confiance de modèle est calculé pour chaque modèle.
Ces scores sont produits dans le fichier de sortie *.cnv.purity.coverage.models.tsv. La fonction de définition de variants somatiques du nombre de copies choisit la paire (pureté, couverture) avec le logarithme de rapport de probabilité le plus élevé.

Si la confiance du modèle choisi est basse, alors la valeur « lowModelConfidence » est inscrite dans le champ FILTER de tous les enregistrements du fichier VCF de sortie.

Fichier VCF de sortie de variants somatiques du nombre de copies

Le fichier VCF de variants somatiques du nombre de copies est en format VCF standard et diffère du fichier VCF de sortie de variants germinaux du nombre de copies des manières suivantes.

Les lignes d'en-tête suivantes sont propres à la définition de variants somatiques du nombre de copies.

- EstimatedTumorPurity (Estimation de la pureté de la tumeur) – L'estimation de la proportion de cellules de l'échantillon résultant d'une tumeur. L'intervalle de valeur de ce champ est [0, 1].
- DiploidCoverage (Couverture diploïde) – Le dénombrement de lectures prévu pour le compartiment de cibles dans une région diploïde. Il n'y a pas de limite à la valeur numérique.
- OverallPloidy (Ploïdie globale) – La moyenne pondérée de la longueur du nombre de copies de tumeur pour les événements portant la marque « PASS » (réussite). Il n'y a pas de limite à la valeur numérique.
- ModelPrecisionScore (Score de précision du modèle) – Le score de confiance propre au modèle qui mesure la distance jusqu'au modèle le plus éloigné pris en considération. L'intervalle de valeur de ce champ est [0, 1].
- ModelAlternativePeakScore (Score du deuxième pic du modèle) – Le score de confiance propre au modèle qui mesure la probabilité d'élimination au deuxième pic le plus élevé de la surface de probabilité. L'intervalle de valeur de ce champ est [0, 1].

La colonne ID représente le type d'événement. En plus de représenter des événements GAIN, LOSS et REF, les entrées CNLOH (perte d'hétérozygotie avec même nombre de copies) et GAINLOH (augmentation du nombre de copies avec perte d'hétérozygotie) représentent des événements LOH (perte d'hétérozygotie).

Le champ ALT peut contenir deux allèles, comme <DEL> et <DUP>, ce qui permet de représenter les nombres de copies propres aux allèles si ceux-ci ont des états de nombre de copies différents.

Le champ FILTER permet d'appliquer les filtres additionnels ci-dessous.

- binCount (Dénombrement de compartiment) – Les événements de variants du nombre de copies dont le dénombrement du compartiment est inférieur au seuil sont filtrés.
- lowModelConfidence (Confiance de modèle basse) – Si la confiance de l'estimation du modèle est basse, tous les enregistrements sont marqués comme non réussis.

Les champs FORMAT sont décrits dans la section de l'en-tête. Les champs suivants sont propres aux variants somatiques du nombre de copies :

- AS – Le nombre de sites de dénombrement de lectures alléliques.

- BC – Le nombre de compartiments de dénombrement de lectures.
- CN – L'estimation du nombre total de copies dans la partie tumorale de l'échantillon.
- CNF – L'estimation à virgule flottante du nombre de copies de tumeur.
- CNQ – Le score de qualité du nombre total exact de copies.
- MAF – L'estimation maximale a posteriori de la fréquence d'allèle mineur.
- MCN – L'estimation du nombre de copies d'haplotype mineur.
- MCNF – L'estimation à virgule flottante du nombre de copies d'haplotype mineur de tumeur.
- MCNQ – Le score de qualité du nombre de copies mineur.
- SD – La meilleure estimation du dénombrement de lectures du segment après la correction du biais.

Détection des expansions de répétition avec la méthode de détection ExpansionHunter

Les petites répétitions en tandem (STR) sont des régions du génome composées de répétitions de petits segments d'ADN appelées unités de répétition. Les petites répétitions en tandem peuvent s'étendre à des longueurs excédant l'étendue habituelle, ce qui entraîne des mutations appelées expansions de répétition. Les expansions de répétition sont responsables de plusieurs maladies, notamment le syndrome de l'X fragile, la sclérose latérale amyotrophique et la maladie de Huntington.

La plateforme DRAGEN comprend une méthode de détection des expansions de répétition appelée ExpansionHunter. Cette méthode effectue un réalignement exact basé sur le graphique de séquences des lectures qui se trouvent dans chaque répétition cible et autour de celle-ci. Elle effectue ensuite le génotypage de la longueur de la répétition sur chaque allèle selon les alignements du graphique. Plus de renseignements et d'analyses sont disponibles dans les articles sur ExpansionHunter suivants :

- ExpansionHunter (version initiale). Consultez la page https://genome.cshlp.org/content/27/11/1895.
- ExpansionHunter avec graphique (nouvelle version intégrée à la plateforme DRAGEN). Consultez la page https://academic.oup.com/bioinformatics/article/35/22/4754/5499079.

Il est important de souligner que ces méthodes ne fonctionnent qu'avec des échantillons de génome humain entier générés avec des méthodes sans PCR. Les répétitions ne sont génotypées que si la couverture du locus est d'au moins 10x.

Options de la fonction de détection des expansions de répétition

Pour activer la détection des expansions de répétition de la plateforme DRAGEN, les options de ligne de commande suivantes sont requises.

- --repeat-genotype-enable = true
- --repeat-genotype-specs = <chemin d'accès vers le fichier de spécifications>

De plus, le sexe associé à l'échantillon doit être spécifié à l'aide de l'option *--sample-sex*.

Les options suivantes sont facultatives.

- --repeat-genotype-region-extension-length = <longueur de la région à examiner autour de la répétition> (la valeur par défaut est 1000 paires de bases)
- --repeat-genotype-min-baseq = <qualité de base minimale pour les bases à fiabilité élevée> (la valeur par défaut est 20)

Pour en savoir plus sur le fichier de spécifications défini par l'option *--repeat-genotype-specs*, consultez la section *Fichiers de spécifications des expansions de répétition* à la page 83.

La sortie principale de la détection des expansions de répétition est un fichier VCF qui contient les variants trouvés grâce à cette analyse.

Fichiers de spécifications des expansions de répétition

Le fichier JSON des spécifications de répétition (également appelé le catalogue des variants de répétitions) décrit les régions de répétition pour l'analyse par la méthode ExpansionHunter. Les spécifications de répétition par défaut pour certaines répétitions pathogènes se trouvent dans le répertoire /opt/edico/repeat-specs/ (selon le génome de référence utilisé avec la plateforme DRAGEN).

Vous pouvez créer des fichiers de spécifications pour les nouvelles régions de répétition en utilisant un des fichiers de spécifications fournis comme modèle. Consultez la documentation relative à la méthode ExpansionHunter pour en savoir plus sur le format.

Fichiers de sortie des expansions de répétition

Fichier VCF de sortie

Les résultats du génotypage des répétitions sont produits dans un fichier VCF distinct dans lequel se trouve la longueur de chaque allèle de toutes les répétitions définissables qui ont été définies dans le fichier de catalogue des spécifications de répétition. Le nom de ce fichier est <outputPrefix>.repeats.vcf (.gz).

Le fichier VCF de sortie commence avec les champs suivants.

Tableau 2 Champs de base du fichier VCF

Champ	Description
CHROM	L'identifiant du chromosome
POS	La position de la première base avant la région de répétition dans la référence
ID	La valeur est toujours « . »
REF	La base de référence à la position POS
ALT	La liste des allèles de répétitions en format <STRn> où n est le nombre d'unités de répétition
QUAL	La valeur est toujours « . »
FILTER	La valeur est toujours « PASS » (réussite)

Tableau 3 Champs INFO additionnels

Champ	Description
SVTYPE	La valeur est toujours « STR »
END	La position de la dernière base de région de la répétition dans la référence
REF	Le nombre d'unités de répétition que la répétition a produit dans la référence
RL	La longueur de la référence en paires de bases
RU	L'orientation de l'unité de répétition dans la référence
REPID	L'identifiant de la répétition dans le fichier de spécifications de répétition

Tableau 4 Champs GENOTYPE (par échantillon)

Champ	Description
GT	Le génotype
SO	Le type de lecture qui contient l'allèle. Sa valeur peut être « SPANNING », « FLANKING » ou « INREPEAT », ce qui signifie respectivement que la lecture englobe une répétition, flanque une répétition ou se trouve au sein d'une répétition.
CI	La longueur de l'intervalle de confiance de la répétition définie pour chaque allèle.
AD_SP	Le nombre de lectures englobant une répétition qui sont cohérentes avec l'allèle.
AD_FL	Le nombre de lectures flanquant une répétition qui sont cohérentes avec l'allèle.
AD_IR	Le nombre de lectures se trouvant au sein d'une répétition qui sont cohérentes avec l'allèle.

Par exemple, l'entrée de fichier VCF suivant décrit l'état de la répétition C9orf72 dans un échantillon avec l'identifiant LP6005616-DNA_A03.

```
QUAL    FILTER  INFO    FORMAT  LP6005616-DNA_A03
chr9    27573526        .       C       <STR2>,<STR349> .       PASS
SVTYPE=STR;END=27573544;REF=3;RL=18;RU=GGCCCC;REPID=ALS GT:SO:CN:CI:AD_SP:AD_
FL:AD_IR
1/2:SPANNING/INREPEAT:2/349:2-2/323-376:19/0:3/6:0/459
```

Dans cet exemple, le premier allèle contient 2 unités de répétition et le deuxième allèle contient 349 unités de répétition. L'unité de répétition est GGCCCC (champ RU de l'entrée INFO). Donc, la séquence du premier allèle est GGCCCCGGCCCC et celle du deuxième est GGCCCC x 349. La répétition contient trois unités de répétition dans la référence (champ REF de l'entrée INFO).

La longueur de l'allèle court a été estimée à l'aide des lectures englobant la répétition (SPANNING), alors que celle de l'allèle étendu a été estimée à l'aide des lectures se trouvant au sein de la répétition (INREPEAT). L'intervalle de confiance pour la taille de l'allèle étendu est (323 à 376). Il y a 19 lectures englobant la répétition et 3 lectures flanquant la répétition qui sont cohérentes avec l'allèle de répétition de taille 2 (c.-à-d. qu'il y a 19 lectures qui contiennent la répétition de taille 2 et 2 lectures flanquant la répétition qui chevauchent tout au plus 2 unités de répétition). Il y a également 6 lectures flanquant de répétition et 459 lectures se trouvant au sein de la répétition qui sont cohérentes avec l'allèle de répétition de taille 349.

Fichiers de sortie supplémentaires

Les alignements du graphique de séquences des lectures des régions de répétition ciblées sont produits dans un fichier BAM. Vous pouvez vous servir d'un outil spécialisé comme GraphAlignmentViewer (github.com/Illumina/GraphAlignmentViewer/) pour visualiser les d'alignements. Les programmes comme Integrative Genomics Viewer (IGV) ne sont pas conçus pour afficher des graphiques de lectures alignées et ne peuvent servir à la visualisation de ces fichiers BAM.

Les fichiers BAM stockent les alignements de graphiques à l'aide de balises XG personnalisées utilisant le format `<LocusName>,<StartPosition>,<GraphCIGAR>`.

- LocusName – Un identifiant de locus identique à l'entrée correspondante dans le fichier de spécifications des expansions de répétition.
- StartPosition – La position d'alignement de départ d'une lecture sur le premier nœud du graphique.
- GraphCIGAR – L'alignement d'une lecture par rapport au graphique ayant cette position de départ. GraphCIGAR est composé d'une séquence d'identifiants de nœuds du graphique et de chaînes CIGAR linéaires qui décrivent l'alignement d'une lecture par rapport à chaque nœud.

Les scores de qualité dans un fichier BAM sont binaires. Un score de 40 est attribué aux bases ayant un score élevé et un score de 0 est attribué aux bases qui ont un score bas.

Fonction de définition de l'amyotrophie spinale

L'interruption de toutes les copies du gène SMN1 d'un individu entraîne l'amyotrophie spinale, une maladie évolutive qui touche environ 1 naissance vivante sur 10 000. Le gène SMN1 a un paralogue à très forte similarité, le gène SMN2, qui n'en diffère que par approximativement 10 VNC et petits indels. Un de ceux-ci (hg19 chr5:70247773 C->T) a un effet sur l'épissage et interrompt la production de la protéine fonctionnelle SMN du gène SMN2. L'analyse de séquençage du génome entier ne produit pas des résultats complets de définition de variants pour la protéine SMN en raison d'une grande similarité des répétitions et de la variation du nombre de copies. Toutefois, environ 95 % des cas d'amyotrophie spinale peuvent être détectés en vérifiant l'absence de l'allèle C fonctionnel (gène SMN1) dans les copies de la protéine SMN.

La plateforme DRAGEN se sert du réalignement de graphique de séquences (tout comme pour la définition des expansions de répétition) pour aligner les lectures à une seule référence représentant les gènes SMN1 et SMN2. En plus du typage génotypique diploïde standard, le programme utilise un test statistique direct pour vérifier la présence d'allèle C. Si aucun allèle C n'est détecté, l'échantillon est défini comme atteint. Sinon, il est défini comme non atteint.

La définition de l'amyotrophie spinale n'est prise en charge que pour les échantillons de séquençage du génome humain entier avec des librairies sans PCR.

Utilisation

La fonction de définition de l'amyotrophie spinale est utilisée conjointement à la fonction de détection des expansions de répétition. Pour en savoir plus sur l'alignement graphique et les options, consultez la section *Détection des expansions de répétition avec la méthode de détection ExpansionHunter* à la page 82.

La fonction de définition de l'amyotrophie spinale, ainsi que la fonction de détection des expansions de répétition, est activée en mettant à l'option *--repeat-genotype-enable* la valeur « true » (activée). Pour activer la fonction de définition de l'amyotrophie spinale, le fichier de catalogue des spécifications de variants doit contenir une description du variant SMN1/2 ciblé. Des fichiers d'exemple sont fournis dans le dossier /opt/edico/repeat-specs/experimental.

Les sorties relatives à la protéine SMN, ainsi que les répétitions ciblées, sont incluses dans le fichier <outputPrefix>.repeat.vcf. La sortie relative à la protéine SMN est représentée par une définition de variants de polymorphisme mononucléotidique à la position clé (celle ayant un effet sur l'épissage) sur le gène SMN1 avec l'état SMA dans les champs personnalisés :

Tableau 5 Résultat relatif à l'amyotrophie spinale dans le fichier de sortie repeat.vcf

Champ	Description
VARID	La valeur « SMN » indique qu'il s'agit d'une définition de protéine SMN.
GT	Le typage génotypique à cette position avec un modèle de génotype de référence (diploïde).
DST	La définition de l'état SMA : « + » indique une détection, « - » indique une absence de détection et « ? » indique une valeur indéterminée.
AD	Le dénombrement total de lectures qui où se trouvent l'allèle C et T.
RPL	Rapport de probabilité en log10 entre les modèles atteint et non atteint. Les scores positifs favorisent le modèle non atteint.

Définition de variants structurels

La fonction de définition de variants structurels de la plateforme DRAGEN intègre et étend les méthodes de la fonction de définition de variants structurels Manta pour fournir des définitions de variants structurels et d'indels de 50 bases ou plus. Les variants structurels et les indels sont définis à partir de lectures de séquençage appariées. La fonction de définition de variants structurels est optimisée pour l'analyse de variations germinales dans des petits ensembles d'individus et des variations somatiques dans des paires d'échantillons de tumeur et de référence.

La fonction de définition de variants structurels effectue les actions suivantes :

- La découverte, l'assemblage et la notation de variants structurels à grande échelle, d'indels de taille moyenne et de grandes insertions au cours d'un seul flux de travail efficace.
- La combinaison des données probantes des lectures appariées et fractionnées pendant la découverte et la notation de variants structurels afin d'améliorer l'exactitude. L'assemblage réussi de lectures fractionnées ou de points de rupture n'est toutefois pas nécessaire pour signaler un variant s'il y a des solides données probantes autres.
- La fonction offre des modèles de notation pour les variants germinaux dans des petits ensembles d'échantillons diploïdes et des variants somatiques dans des paires d'échantillons de tumeur et de référence.

Une prise en charge expérimentale est aussi offerte aux fins d'analyse des échantillons de tumeur sans correspondance. Toutes les inférences de variants structurels et d'indels sont produites dans le format VCF 4.1.

Supprimez ce texte et remplacez-le par votre propre contenu.

Aperçu de la fonction de définition de variants structurels de la plateforme DRAGEN

La fonction de définition de variants structurels de la plateforme DRAGEN divise le processus de découverte de variants structurels et d'indels en deux grandes étapes :

1 Le balayage du génome permet de construire différentes structures de données du génome entier, y compris un graphique d'association des extrémités rompues pour toutes les régions associées à des variants structurels. Le graphique montre les périphéries reliant les régions du génome qui pourraient être associées par l'intermédiaire d'extrémités rompues. Les périphéries peuvent connecter deux régions différentes du génome pour constituer des données probantes d'une association à distance. Une périphérie peut également connecter une région à elle-même pour indiquer une association locale d'indel ou de petit variant structurel. Ces associations sont plus générales qu'une certaine hypothèse de variant structurel et plusieurs extrémités rompues candidates peuvent se trouver sur une même périphérie. Généralement, il n'y a qu'un ou deux candidats par périphérie.

2 L'analyse des périphéries ou des groupes de périphéries très connectées du graphique permet de déceler et de noter les variants structurels associés aux périphéries. La fonction de définition de variants structurels effectue son analyse comme suit :

   a Inférence de variants structurels candidats associés à la périphérie.
   b Tentative d'assemblage des extrémités rompues des variants structurels.
   c Notation, génotypage et filtre du variant structurel selon différents modèles biologiques (modèles diploïde, germinal et somatique actuellement).
   d Sortie en fichier VCF.

Capacités de la fonction de définition de variants structurels de la plateforme DRAGEN

La fonction de définition de variants structurels de la plateforme DRAGEN peut détecter tous les types de variants structurels qui sont identifiables en l'absence d'analyse du nombre de copies et d'assemblage de novo à grande échelle. Pour en savoir plus à ce sujet, consultez la section *Classes de variants détectés* à la page 87.

Pour chaque variant structurel et indel, la fonction de définition de variants structurels tente d'assembler les extrémités rompues à une résolution au niveau des paires de bases et de signaler les coordonnées de l'extrémité rompue déplacée vers la gauche (conformément aux recommandations pour les rapports VCF de variants structurels 4.1), avec toutes les séquences homologiques ou insérées entre les extrémités rompues. Il arrive souvent que l'assemblage ne réussisse pas à fournir une explication convaincante des données. Si c'est le cas, le variant est signalé comme IMPRECISE (IMPRÉCIS) et sera seulement noté selon les données probantes de la lecture appariée.

Il est attendu que les lectures de séquençage fournies en entrée à la fonction de définition de variants structurels proviennent d'un test de séquençage de lectures appariées. Ceci entraîne une orientation vers l'intérieur entre les deux lectures de chaque fragment de séquence, car chacune a une lecture de la périphérie de l'insert du fragment orientée vers son centre.

La fonction de définition de variants structurels est testée principalement pour les tests de séquençage de génome entier et d'exome entier (ou autre enrichissement de cibles) sur l'ADN. Pour ces tests, les applications suivantes sont prises en charge :

- Analyse combinée de petits ensembles d'individus diploïdes (où petit signifie à l'échelle de la famille, c'est-à-dire 10 échantillons ou moins)
- Analyse soustractive d'une paire d'échantillons de tumeur et de référence correspondants
- Analyse d'un seul échantillon de tumeur

Pour l'analyse combinée, il n'y a pas de restrictions particulières par rapport à l'utilisation de la fonction de définition de variants structurels pour l'analyse combinée de grandes cohortes. Comme son utilisation n'a pas fait l'objet d'un examen approfondi, des problèmes de stabilité ou de qualité de définition pourraient survenir.

Les échantillons de tumeur peuvent être analysés sans échantillon de référence correspondant. Dans ce cas, aucune fonction de notation n'est disponible, mais les dénombrements de données probantes à l'appui et plusieurs filtres peuvent quand même être utiles.

Classes de variants détectés

La fonction de définition de variants structurels détecte toutes les classes de variations qui peuvent être définies comme de nouvelles adjacences d'ADN dans le génome. Toutes les nouvelles adjacences d'ADN peuvent être classées dans les catégories suivantes selon la structure de l'extrémité rompue :

- Délétions
- Insertions
  - Insertions complètement assemblées
  - Insertions partiellement assemblées (c.-à-d. inférées)
- Répétitions en tandem
- Paires d'extrémités rompues correspondantes à des translocations intra et interchromosomiques ou à des variants structurels complexes.

Limites connues

La fonction de définition de variants structurels ne devrait pas être en mesure de détecter les types de variants suivants :

- Les répétitions dispersées.
- La plupart des variants d'expansion et de contraction d'une répétition en tandem de référence.
- Les extrémités rompues qui correspondent à de petites inversions.
  - La taille limite n'a pas été testée. En théorie, il n'y a pas de détection sous environ 200 bases. Les supposées micro-inversions peuvent être indirectement détectées en tant que variants d'insertion et de délétion.
- Les grandes insertions complètement assemblées.
  - La taille maximale des insertions complètement assemblées devrait être d'environ deux fois la taille du fragment de la paire de lectures. La capacité d'assemblage complet de l'insertion devrait devenir pratiquement irréalisable en deçà de cette taille.
  - La fonction de définition de variants structurels peut détecter et signaler de très grandes insertions quand la signature d'extrémité rompue d'un tel événement est décelée et ce, même si la séquence insérée ne peut être complètement assemblée.

Il y a également d'autres limites générales liées aux répétitions pour tous les types de variants :

- La capacité à assembler un variant à la résolution de l'extrémité rompue devient nulle quand la longueur de la répétition de l'extrémité rompue est près de la taille de la lecture.
- La capacité à détecter toute extrémité rompue devient (presque) nulle quand la longueur de la répétition de l'extrémité rompue est près de la taille du fragment.

Bien que la fonction de définition de variants structurels soit en mesure de classifier de nouvelles adjacences d'ADN, elle n'infère pas les structures de plus haut niveau sous-entendues par la classification. Par exemple, un variant marqué comme délétion par la fonction de définition de variants structurels indique une translocation intrachromosomique avec une structure d'extrémité rompue qui correspond à une délétion. Il n'y a toutefois pas d'examen de la profondeur, de la fréquence de l'allèle b ou des adjacences qui se recoupent pour inférer directement le type de variants structurels.

Exigences d'entrée

Quand vous utilisez la fonction de définition de variants structurels en mode autonome, il est attendu que les lectures de séquençage fournies en entrée proviennent d'un test de séquençage à lectures appariées avec une orientation vers l'intérieur entre les deux lectures de chaque fragment d'ADN, car chacun a une lecture de la périphérie de l'insert de fragment orientée vers son centre.

La fonction de définition de variants structurels peut tolérer des lectures non appariées en entrée tant qu'il y a suffisamment de lectures appariées pour estimer la distribution de taille des fragments appariés. Les lectures non appariées sont tout de même utilisées pour la découverte, l'assemblage et la notation de lectures fractionnées si leurs alignements (ou leurs alignements fractionnés avec étiquette SA) prend en charge un indel ou un variant structurel de grande taille ou si un mésappariement ou un découpage suggère qu'il s'agit de l'emplacement possible d'une extrémité rompue.

En mode autonome, les lectures de séquençage d'entrée doivent être cartographiées et fournies comme entrée au format BAM ou CRAM. Chaque fichier d'entrée doit être trié par coordonnées et indexé pour produire un index de style samtools/htslib dans un fichier au nom correspondant au fichier d'entrée BAM ou CRAM auquel est ajouté l'extension .bai, .crai ou .csi.

Au moment de la configuration, au moins un fichier BAM ou CRAM doit être fourni pour un échantillon de tumeur ou de référence. Une paire d'échantillons de tumeur et de référence correspondants peut également être fournie. Si plusieurs fichiers d'entrée sont fournis pour l'échantillon de référence, chaque fichier est traité comme un échantillon séparé dans le cadre d'une analyse combinée d'échantillons diploïdes.

En mode autonome, les fichiers d'entrée BAM et CRAM ont les restrictions suivantes :

- Les alignements ne peuvent pas avoir de séquence de lecture inconnue (champ SEQ contenant le caractère « * »).
- Le champ SEQ des alignements ne peut pas contenir le caractère « = ».
- Les alignements ne peuvent pas utiliser l'appariement ou le mésappariement de séquence (caractères « = » et « X »). Les étiquettes de groupes de lectures de la notation CIGAR qui se trouvent dans les enregistrements d'alignement sont ignorées. Chaque alignement est traité comme s'il représentait un échantillon.
- Les alignements ayant des valeurs de qualité de définition de bases supérieures à 70 sont rejetés (ils ne sont pas pris en charge, car on suppose que cela indique une erreur de décalage).

Options relatives à la définition de variants structurels

Les options de ligne de commande suivantes sont prises en charge pour la fonction de définition de variants structurels :

- *--enable-sv* – Permet d'activer ou de désactiver la fonction de définition de variants structurels. La valeur par défaut est « false » (désactivée).
- *--sv-call-regions-bed* – Spécifie un fichier BED contenant un ensemble de régions à définir. Le fichier peut également être compressé en format gzip ou bgzip.
- *--sv-region* – Limite l'analyse à une région spécifiée du génome aux fins de débogage. Cette option peut être spécifiée plusieurs fois pour créer une liste de régions. La valeur doit être dans le format « chr:startPos-endPos ».
- *--sv-exome* – Lorsque la valeur est « true » (activée), l'option configure la fonction de définition de variants pour les entrées de séquençage ciblées, y compris en désactivant les filtres de grandes profondeurs. La valeur par défaut est « false » (désactivée).
- *--sv-output-contigs* – La valeur « true » (activée) permet de produire des séquences de contigs assemblées dans un fichier VCF. La valeur par défaut est « false » (désactivée).

Modes de fonctionnement

La définition de variants structurels peut être exécutée dans les modes suivants :

- Autonome – Mode qui utilise les fichiers BAM/CRAM cartographiés en entrée. Ce mode requiert les options suivantes :
  - *--enable-map-align false*
  - *--enable-sv true*
- Intégré – Mode exécuté automatiquement sur la sortie de la fonction de cartographie et d'alignement de la plateforme DRAGEN. Ce mode requiert les options suivantes :
  - *--enable-map-align true*
  - *--enable-sv true*
  - *--enable-map-align-output true*
  - *--output-format bam*

La définition de variants structurels peut également être activée avec toute autre fonction de définition.

Voici un exemple de ligne de commande pour le mode intégré :

```
dragen -f \
    --ref-dir=<HASH_TABLE> \
   --enable-map-align true \
   --enable-map-align-output true \
   --enable-sv true \
   --output-directory <OUT_DIR> \
   --output-file-prefix <PREFIX> \
   --RGID Illumina_RGID \
   --RGSM <sample name> \
   -1 <FASTQ1> \
   -2 <FASTQ2>
```

Voici un exemple de ligne de commande pour la définition diploïde combinée en mode autonome :

```
dragen -f \
   --ref-dir <HASH_TABLE> \
   --bam-input <BAM1> \
   --bam-input <BAM2> \
   --bam-input <BAM3> \
   --enable-map-align false \
   --enable-sv true \
   --output-directory <OUT_DIR> \
   --output-file-prefix <PREFIX>
```

Fichier VCF de sortie de variants structurels

Le fichier VCF de sortie de variants structurels est disponible dans le répertoire de sortie. Le fichier est nommé *<output-file-prefix>.sv.vcf.gz*.

La fonction de définition de variants structurels produit des fichiers de sortie additionnels dans le répertoire<output-directory>/sv/. Le dossier <output-directory>/sv/results contient des fichiers de sortie de statistiques et de variants additionnels. Les sous-répertoires additionnels contiennent des journaux et des sorties intermédiaires du processus de définition de variants.

Prédictions de variants structurels

Dans le répertoire <output-directory>/sv/results/variants, la fonction de définition de variants structurels produit un ensemble de fichiers VCF 4.1. Actuellement, deux fichiers VCF sont créés pour une analyse germinale et un fichier VCF de variants somatiques additionnel est produit pour effectuer une analyse soustractive d'échantillons de tumeur et de référence. Ces fichiers sont les suivants :

- diploidSV.vcf.gz

  Les variants structurels et les indels notés et génotypés selon un modèle diploïde pour un ensemble d'échantillons dans une analyse combinée d'échantillons diploïdes ou pour l'échantillon de référence d'une analyse soustractive d'échantillons de tumeur et de référence. Dans le cas d'une analyse soustractive d'échantillons de tumeur et de référence, les scores de ce fichier ne révèlent aucun renseignement relatif à l'échantillon de tumeur.

- somaticSV.vcf.gz

  Les variants structurels et les indels sont notés selon un modèle de variants somatiques. Ce fichier n'est produit que si le fichier d'alignement de l'échantillon de tumeur est fourni pendant la configuration.

- candidateSV.vcf.gz

  Les candidats de variants structurels et d'indels non notés. Très peu de données probantes à l'appui sont nécessaires pour qu'un variant structurel soit ajouté comme candidat dans ce fichier. Un variant structurel ou un indel doit être un candidat avant d'être noté. Un variant structurel ne peut donc apparaître dans les autres fichiers VCF de sortie s'il ne se trouve pas d'abord dans ce fichier. Ce fichier contient les indels d'une taille de 8 et plus. Les plus petits indels sont fournis pour donner de la flexibilité au flux de travail, mais ne sont pas notés. Seuls les variants structurels d'une taille de 50 et plus sont notés.

Pour l'analyse d'échantillon de tumeur seulement, la fonction de définition de variants structurels produit les fichiers VCF de sortie additionnels suivants :

- tumorSV.vcf.gz

  Un sous-ensemble du fichier candidateSV.vcf.gz obtenu après avoir retiré les candidats redondants et les petits indels dont la taille est inférieure au seuil de notation (50). Les variants structurels ne sont pas notés, mais les détails additionnels suivants sont inclus : (1) les dénombrements de données probantes à l'appui des lectures appariées et fractionnées pour chaque allèle, (2) un sous-ensemble des filtres du modèle d'échantillons de tumeur et de référence sont appliqués au modèle à un échantillon de tumeur pour améliorer la précision.

Fichier VCF de sortie

Le fichier VCF de sortie répond à la spécification VCF 4.1 pour la description des variants structurels. Il utilise des noms de champ standards dans la mesure du possible. Tous les champs personnalisés sont décrits dans l'en-tête du fichier VCF. Les sections qui suivent contiennent les détails de la représentation des variants et les principales valeurs des champs du fichier VCF.

Noms d'échantillon de fichier VCF

Les noms d'échantillon produits dans le fichier VCF de sortie sont extraits de chaque fichier d'alignement en entrée à partir du premier enregistrement de groupe de lectures (@RG) trouvé dans l'en-tête. Tous les espaces se trouvant dans le nom sont remplacés par des traits de soulignement. Si aucun nom d'échantillon n'est trouvé, une étiquette par défaut (SAMPLE1, SAMPLE2, etc.) est utilisée.

Petit indel

Tous les variants sont signalés dans le fichier VCF à l'aide d'allèles symboliques sauf s'ils sont classés en tant que petits indels. Dans ce cas, les séquences complètes sont fournies dans les champs d'allèles REF et ALT du fichier VCF. Un variant est classé en tant que petit indel si tous les critères suivants sont réunis :

- Le variant peut être complètement exprimé par une combinaison de séquences d'insertions et de délétions.
- La longueur de la délétion ou de l'insertion est inférieure à 1000.
- Les extrémités rompues du variant et/ou la séquence d'insertion ne sont imprécises.

Quand les enregistrements du fichier VCF sont produits en sortie au format de petits indels, ils comprennent également l'étiquette CIGAR INFO qui décrit l'événement combiné d'insertion et de délétion.

Insertions avec assemblage de séquence d'insert incomplet

Les grandes insertions sont parfois signalées même quand la séquence d'insert ne peut être complètement assemblée. Dans ce cas, la fonction de définition de variants structurels signale l'insertion avec l'allèle symbolique <INS> et ajoute les champs INFO spéciaux LEFT_SVINSSEQ et RIGHT_SVINSSEQ afin de décrire les extrémités assemblées de gauche et de droite de la séquence d'insert. Ci-dessous se trouve un exemple d'un tel enregistrement provenant de l'analyse combinée de diploïdes des échantillons NA12878, NA12891 et NA12892 cartographiés au hg19 :

```
chr1 11830208 MantaINS:1577:0:0:0:3:0 T <INS> 999 PASS
END=11830208;SVTYPE=INS;CIPOS=0,12;CIEND=0,12;HOMLEN=12;HOMSEQ=TAAATTTT
TCTT;LEFT_
SVINSSEQ=TAAATTTTTCTTTTTTCTTTTTTTTTTAAATTTATTTTTTTATTGATAATTCTTGGGTGTTT
CTCACAGAGGGGGATTTGGCAGGGTCACGGGACAACAGTGGAGGGAAGGTCAGCAGACAAACAAGTGAACA
AAGGTCTCTGGTTTTCCCAGGCAGAGGACCCTGCGGCCTTCCGCAGTGTTCGTGTCCCTGATTACCTGAGA
TTAGGGATTTGTGATGACTCCCAACGAGCATGCTGCCTTCAAGCATCTGTTCAACAAAGCACATCTTGCAC
TGCCCTTAATTCATTTAACCCCGAGTGGACACAGCACATGTTTCAAAGAG;RIGHT_
SVINSSEQ=GGGGCAGAGGCGCTCCCCACATCTCAGATGATGGGCGGCCAGGCAGAGACGCTCCTCACTTC
CTAGATGTGATGGCGGCTGGGAAGAGGCGCTCCTCACTTCCTAGATGGGACGGCGGCCGGGCGGAGACGCT
CCTCACTTTCCAGACTGGGCAGCCAGGCAGAGGGGCTCCTCACATCCCAGACGATGGGCGGCCAGGCAGAG
ACACTCCCCACTTCCCAGACGGGGTGGCGGCCGGGCAGAGGCTGCAATCTCGGCACTTTGGGAGGCCAAGG
CAGGCGGCTGCTCCTTGCCCTCGGGCCCCGCGGGGCCCGTCCGCTCCTCCAGCCGCTGCCTCC
GT:FT:GQ:PL:PR:SR 0/1:PASS:999:999,0,999:22,24:22,32
0/1:PASS:999:999,0,999:18,25:24,20 0/0:PASS:230:0,180,999:39,0:34,0
```

Inversions

Par défaut, les inversions sont signalées en tant qu'extrémités rompues. Pour une inversion réciproque simple, quatre extrémités rompues sont signalées et elles ont toute la même étiquette EVENT INFO. Voici un exemple d'inversion réciproque simple :

```
chr1 17124941 MantaBND:1445:0:1:1:3:0:0 T [chr1:234919886[T 999 PASS
SVTYPE=BND;MATEID=MantaBND:1445:0:1:1:3:0:1;CIPOS=0,1;HOMLEN=1;
HOMSEQ=T;INV5;EVENT=MantaBND:1445:0:1:0:0:0:0;JUNCTION_QUAL=254;BND_
DEPTH=107;
MATE_BND_DEPTH=100 GT:FT:GQ:PL:PR:SR 0/1:PASS:999:999,0,999:65,8:15,51
chr1 17124948 MantaBND:1445:0:1:0:0:0:0 T T]chr1:234919824] 999 PASS
SVTYPE=BND;MATEID=MantaBND:1445:0:1:0:0:0:1;INV3;EVENT=MantaBND:1445:0:
1:0:0:0:0;
JUNCTION_QUAL=999;BND_DEPTH=109;MATE_BND_DEPTH=83 GT:FT:GQ:PL:PR:SR
0/1:PASS:999:999,0,999:60,2:0,46
chr1 234919824 MantaBND:1445:0:1:0:0:0:1 G G]chr1:17124948] 999 PASS
SVTYPE=BND;MATEID=MantaBND:1445:0:1:0:0:0:0;INV3;EVENT=MantaBND:1445:0:
1:0:0:0:0;
JUNCTION_QUAL=999;BND_DEPTH=83;MATE_BND_DEPTH=109 GT:FT:GQ:PL:PR:SR
0/1:PASS:999:999,0,999:60,2:0,46
chr1 234919885 MantaBND:1445:0:1:1:3:0:1 A [chr1:17124942[A 999 PASS
SVTYPE=BND;MATEID=MantaBND:1445:0:1:1:3:0:0;CIPOS=0,1;HOMLEN=1;
HOMSEQ=A;INV5;EVENT=MantaBND:1445:0:1:0:0:0:0;JUNCTION_QUAL=254;BND_
DEPTH=100;
MATE_BND_DEPTH=107 GT:FT:GQ:PL:PR:SR 0/1:PASS:999:999,0,999:65,8:15,51
```

Champs de l'entrée INFO de fichier VCF

Identifiant	Description
IMPRECISE	La marque indiquant que la variation structurelle est imprécise. Par exemple, si l'emplacement exact de l'extrémité rompue est inconnu.
SVTYPE	Le type de variant structurel.
SVLEN	La différence de longueur entre les allèles REF et ALT.
END	La position finale du variant décrit dans cet enregistrement.
CIPOS	L'intervalle de confiance par rapport à la valeur du champ POS.
CIEND	L'intervalle de confiance par rapport à la valeur du champ END.
CIGAR	L'alignement CIGAR pour chaque allèle d'indel alternatif.
MATEID	L'identifiant de couple de l'extrémité rompue.
EVENT	L'identifiant de l'événement associé à l'extrémité rompue.
HOMLEN	La longueur de l'homologie identique de paires de bases aux extrémités rompues de l'événement.
HOMSEQ	La séquence de l'homologie identique de paires de bases aux extrémités rompues de l'événement.
SVINSLEN	La longueur de l'insertion.
SVINSSEQ	La séquence de l'insertion.
LEFT_SVINSSEQ	Le côté gauche connu d'une insertion de longueur inconnue.
RIGHT_SVINSSEQ	Le côté droit connu d'une insertion de longueur inconnue.
BND_DEPTH	La profondeur de lecture à l'extrémité rompue de la translocation locale.
MATE_BND_DEPTH	La profondeur de lecture à l'extrémité rompue du couple de la translocation à distance.
JUNCTION_QUAL	Si la jonction du variant structurel fait partie d'un événement (c.-à-d. un variant à plusieurs adjacences), ce champ fournit la valeur QUAL de l'adjacence en question seulement.
SOMATIC	La marque indiquant un variant somatique.
SOMATICSCORE	Le score de qualité du variant somatique.
JUNCTION_ SOMATICSCORE	Si la jonction du variant structurel fait partie d'un événement (c.-à-d. un variant à plusieurs adjacences), ce champ fournit la valeur SOMATICSCORE de l'adjacence en question seulement.
CONTIG	La séquence de contig assemblée, si le variant n'est pas imprécis (avec l'option *--outputContig*).

Champs de l'entrée FORMAT du fichier VCF

Identifiant	Description
GT	Le génotype
FT	Le filtre d'échantillon. La valeur « PASS » signifie que cet échantillon a passé tous les filtres (réussite).
GQ	La qualité du génotype
PL	Les scores normalisés de probabilité de l'échelle Phred pour les génotypes tels que définis dans la spécification de fichier VCF.
PR	Le nombre de paires de lectures contenant toute la répétition qui soutiennent fortement (Q30) les allèles REF ou ALT
SR	Le nombre de paires de lectures fractionnées qui soutiennent fortement (Q30) les allèles REF ou ALT

Champs FILTER du fichier VCF

Identifiant	Niveau	Description
MinQUAL	Enregistrement	Score QUAL inférieur à 20
MinGQ	Échantillon	Score GQ inférieur à 15
MinSomaticScore	Enregistrement	SOMATICSCORE inférieur à 30
Ploidy	Enregistrement	Pour les variants DEL et DUP, les génotypes des variants (de tailles similaires) qui se chevauchent ne correspondent pas à une attente diploïdique.
MaxDepth	Enregistrement	La profondeur est supérieure à trois fois la profondeur chromosomique médiane près de l'une ou des deux extrémités rompues du variant.
MaxMQ0Frac	Enregistrement	Pour les petits variants (moins de 1000 bases), la proportion des lectures de tous les échantillons dont une des extrémités rompues a un score MAPQ0 est supérieure à 0,4.
NoPairSupport	Enregistrement	Pour les variants dont la taille est largement supérieure à celle du fragment de la lecture appariée, aucune lecture appariée ne soutient l'allèle alternatif dans tous les échantillons.
SampleFT	Enregistrement	Aucun échantillon ne passe tous les filtres au niveau de l'échantillon.
HomRef	Échantillon	Définition de référence homozygote

Interprétation des fichiers VCF

Il y a deux niveaux de filtres : au niveau de l'enregistrement (FILTER) et au niveau de l'échantillon (FORMAT/FT). Les filtres du niveau de l'enregistrement sont généralement indépendants de ceux du niveau de l'échantillon. Cependant, si aucun des échantillons ne passe tous les filtres du niveau de l'échantillon, le filtre SampleFT est appliqué au niveau de l'enregistrement.

Interprétation du champ INFO/EVENT

Certains variants structurels signalés dans le fichier VCF, comme les translocations, représentent une seule nouvelle jonction de séquence dans l'échantillon. Le champ INFO/EVENT indique qu'au moins deux jonctions de ce type pourraient se produire dans le cadre d'un seul événement de variant. Tous les enregistrements de variant individuel qui font partie du même événement ont la même chaîne INFO/EVENT. Toutefois, bien qu'une telle inférence pourrait être appliquée après la définition de variants structurels en analysant la distance relative et l'orientation des extrémités rompues du variant défini, la fonction de définition de variants structurels inclut ce mécanisme d'événement au processus de définition afin d'améliorer la sensibilité de la détection de tels événements à grande échelle. Comme au moins une jonction de l'événement a déjà passé les seuils standards pour être considérée comme un variant candidat, la sensibilité est améliorée en diminuant les seuils relatifs aux données probantes pour les jonctions additionnelles qui se produisent selon un modèle conséquent à un événement multijonctions (comme une paire de translocations réciproques).

Même si ce mécanisme pourrait être généralisé à des événements qui comprennent un nombre arbitraire de jonctions, il est actuellement limité à ceux qui en ont deux. Donc, pour le moment, sa principale utilité est l'identification des paires de translocations réciproques et l'amélioration de la sensibilité pour les déceler.

Champ ID du fichier VCF

Le champ ID, ou identifiant, du fichier VCF peut servir à l'annotation. Dans les enregistrements BND (extrémités rompues) des translocations, la valeur du champ ID sert à lier les couples ou les partenaires d'extrémités rompues. Voici un exemple de champ ID de fichier VCF.

```
MantaINS:1577:0:0:0:3:0
```

La valeur fournie dans le champ ID reflète la ou les périphéries du graphique d'association de variants structurels où le variant structurel ou l'indel a été décelé. Les détails de la valeur ID fournie sont principalement destinés à une utilisation interne par les développeurs. La valeur est unique au sein d'un fichier VCF de sortie produit par la fonction de définition de variants structurels. Ces valeurs d'identifiant servent à lier les enregistrements des extrémités rompues associées à l'aide de la clé MATEID standard de fichiers VCF. La structure de cet identifiant pourrait être modifiée dans le futur. Vous pouvez vous servir de la valeur entière comme clé unique. Par contre, la décomposer pour en utiliser les différents éléments pourrait entraîner une incompatibilité avec de futures mises à jour.

Conversion de fichier VCF de variants structurels au format BEDPE

Il est parfois pratique d'exprimer les variants structurels en format BEDPE. Pour ce faire, nous recommandons le script vcfToBedpe disponible à partir de la page :

```
https://github.com/ctsa/svtools
```

Il s'agit d'une fourche de référence de @hall-lab présentant des modifications visant la prise en charge du format VCF 4.1 de variants structurels.

Le format BEDPE contient beaucoup moins de renseignements sur les variants structurels que le fichier de sortie VCF de la fonction de définition de variants structurels. En particulier, les orientations des extrémités rompues, l'homologie des extrémités rompues et les séquences d'insertion sont perdues, en plus de la capacité à définir des champs pour les renseignements propres au locus et à l'échantillon. Illumina recommande donc d'utiliser le format BEDPE uniquement comme sortie temporaire pour les applications qui en ont besoin.

Fichier de sortie de statistiques

Les sorties secondaires additionnelles sont fournies dans les fichiers du dossier <output-directory>/sv/results/stats.

- alignmentStatsSummary.txt

  Les quantiles de longueur de fragment pour chaque fichier d'alignement en entrée.

- svLocusGraphStats.tsv

  Les statistiques et les renseignements sur la durée d'exécution relatifs au graphique des locus de variants structurels.

- svCandidateGenerationStats.tsv

  Les statistiques et les renseignements sur la durée d'exécution en lien avec la génération de candidats de variants structurels.

- svCandidateGenerationStats.xml

  Les données XML justificatives du rapport svCandidateGenerationStats.tsv.

- diploidSV.sv_metrics.csv

  Le nombre de définitions de variants structurels réussies avec un modèle diploïde. Ce fichier n'est produit que lors de l'utilisation de l'analyse germinale ou d'échantillons de tumeur et de référence.

- somaticSV.sv_metrics.csv

  Le nombre de définitions de variants structurels réussies avec un modèle de variants somatiques. Ce fichier n'est produit que lors de l'utilisation de l'analyse d'échantillons de tumeur et de référence.

- tumorSV.sv_metrics.csv

  Le nombre de définitions de variants structurels réussies de l'analyse d'échantillons de tumeur et de référence. Ce fichier n'est produit que lors de l'utilisation de l'analyse d'échantillon de tumeur seulement.

Score de qualité de novo pour variants structurels

Un score de qualité de novo peut être activé pour la définition diploïde combinée de variants structurels en mettant à l'option *--sv-denovo-scoring* la valeur « true » (activée) et en fournissant un fichier de généalogie. Cette option ajoute les champs FORMAT/DQ et FORMAT/DN au fichier VCF de sortie pour représenter un score de qualité de novo associé à une définition de novo.

L'exemple suivant montre une ligne de commande permettant d'activer le score de qualité de novo pour une analyse diploïde combinée.

```
dragen -f
   --ref-dir <HASH_TABLE> \
   --bam-input <BAM1> \
   --bam-input <BAM2> \
   --bam-input <BAM3> \
   --enable-map align=false \
   --enable-sv=true \
   --output-directory <OUT_DIR> \
   --output-file-prefix <PREFIX> \
   --sv-denovo-scoring true \
   --RGID DRAGEN_RGID \
   --RGSM <sample name>
   --pedigree-file <PED_FILE>
```

La plateforme DRAGEN peut aussi analyser un fichier VCF de sortie existant pour un variant structurel multiéchantillons (c.-à-d. un trio comprenant un proposant et ses parents) pour générer un fichier VCF modifié contenant des champs FORMAT/DQ et FORMAT/DN (le fichier original reste inchangé).

L'exemple suivant montre une ligne de commande permettant de dériver le score de qualité de novo à partir d'un trio existant de variants structurels.

```
dragen -f \
   --variant <TRIO_VCF_FILE> \
   --pedigree-file <PED_FILE> \
   --enable-map-align false \
   --sv-denovo-scoring true \
   --output-directory <OUT_DIR> \
   --output-file-prefix <PREFIX>
```

Le champ DQ est défini comme suit :

```
##FORMAT=<ID=DQ,Number=1,Type=Float,Description="Denovo quality">
```

Le champ DQ représente un score de probabilité a posteriori que le variant soit de novo chez le proposant. S'il peut être calculé, le score de l'échelle Phred est ajouté au proposant, alors que les autres échantillons sont marqués d'un point (« . ») pour indiquer qu'il est manquant.

Par exemple, des scores DQ de « 13 » et « 20 » correspondraient à une probabilité a posteriori d'un variant de novo de « 0,95 » et « 0,99 », respectivement.

Le champ DN est défini comme suit :

```
##FORMAT=<ID=DN,Number=1,Type=String,Description="Possible values are
   'DeNovo' or 'LowDQ'. Threshold for a passing de novo call is DQ >=
   20">
```

La plateforme DRAGEN compare les scores DQ valides (> 0) avec le seuil d'un score de « 20 » par défaut. Un score supérieur ou égal au seuil se traduit en une mention « de novo » dans le champ DN de l'échantillon, alors qu'un score inférieur au seuil se traduit plutôt en une mention « LowDQ ». Si aucun score DQ n'est valide (c.-à-d. le score DQ est nul ou absent), alors la valeur « . » sera mise dans le champ.

Le seuil peut être modifié à l'aide de l'option de ligne de commande *--sv-denovo-threshold*. Par exemple, pour réduire le seuil à « 10 », ajoutez *--sv-denovo-threshold 10* à la ligne de commande de la plateforme DRAGEN.

Les fichiers d'entrée de cette fonction sont le fichier VCF et le fichier de généalogie qui précise quels échantillons du trio correspondent à celui du proposant, à celui de la mère et à celui du père. Dans le cas où plusieurs trios sont précisés dans le fichier de généalogie (p. ex., une généalogie sur plusieurs générations), la plateforme DRAGEN détecte automatiquement les trios et évalue les variants de novo sur l'échantillon du proposant de chaque trio.

Fonction d'estimation de la ploïdie

La fonction d'estimation de la ploïdie se sert des renseignements de la lecture produits par la fonction de cartographie et d'alignement pour établir un caryotype sexuel en fonction de la couverture du génome.

La fonction d'estimation de la ploïdie est activée par défaut. Vous pouvez activer ou désactiver la fonction d'estimation de la ploïdie grâce à l'option de ligne de commande *--auto-detect-sample-sex*. Pour désactiver la fonction d'estimation de la ploïdie, utilisez la commande suivante :

```
--auto-detect-sample-sex=false
```

La fonction d'estimation de la ploïdie estime les caryotypes sexuels en divisant les couvertures médianes des chromosomes sexuels par la couverture autosomique médiane. Quand ces rapports se trouvent dans certains intervalles, les caryotypes sexuels suivants peuvent être prédits : XX, XY, X0, XXY, XYY, XXX et XXXY.

Caryotype sexuel	Rapport minimal de X	Rapport maximal de X	Rapport minimal de Y	Rapport maximal de Y
XX	0,75	1,25	0,00	0,25
XY	0,25	0,75	0,25	0,75
XXY	0,75	1,25	0,25	0,75
XYY	0,25	0,75	0,75	1,25
X0	0,25	0,75	0,00	0,25
XXXY	1,25	1,75	0,25	0,75
XXX	1,25	1,75	0,00	0,25

Si la fonction d'estimation de la ploïdie réussit à prédire un caryotype sexuel, les résultats sont propagés en aval aux fonctions de définition de petits variants et de génotypage de répétitions. De plus, chaque couverture médiane par contig normalisée est ajoutée dans le fichier __.ploidy_estimation_metrics.csv et dans le fichier de sortie standard. L'exemple suivant montre des résultats en format CSV et dans le fichier de sortie standard.

PLOIDY ESTIMATION	Autosomal median coverage	44.79
PLOIDY ESTIMATION	X median coverage	42.47
PLOIDY ESTIMATION	Y median coverage	20.82
PLOIDY ESTIMATION	1 median/Autosomal median	0.95
PLOIDY ESTIMATION	2 median/Autosomal median	1.05
PLOIDY ESTIMATION	3 median/Autosomal median	1.01
PLOIDY ESTIMATION	4 median/Autosomal median	0.99
…		
PLOIDY ESTIMATION	22 median/Autosomal median	0.99
PLOIDY ESTIMATION	X median/Autosomal median	0.95
PLOIDY ESTIMATION	Y median/Autosomal median	0.46
PLOIDY ESTIMATION	Ploidy estimation	XXY

La fonction d'estimation de la ploïdie peut échouer si le type de données d'entrée (génome ou exome) ne peut être prédit, s'il n'y a pas suffisamment de couverture pour calculer des statistiques ou si les rapports X/Y ne se trouvent pas dans des intervalles connus.

Indicateurs de contrôle de la qualité et rapports sur la couverture et la définissabilité

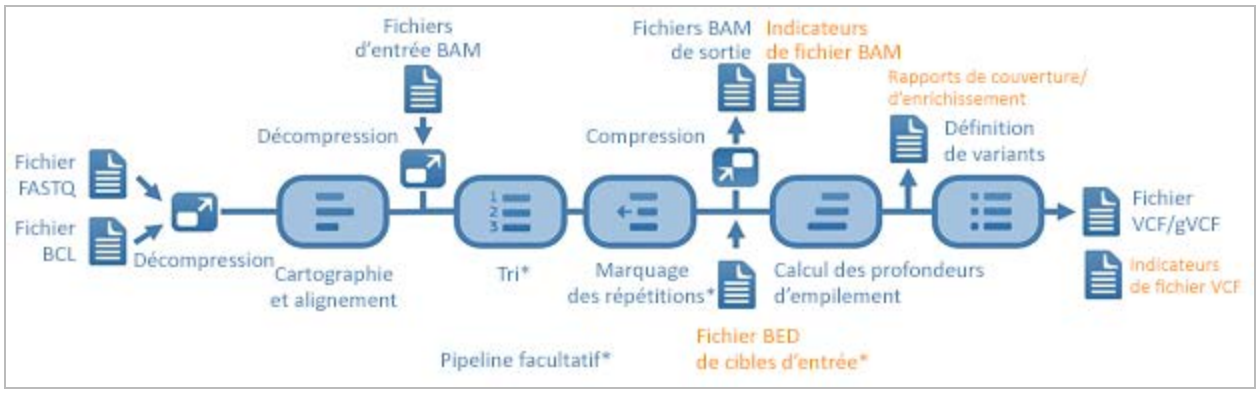
Les rapports sur les indicateurs de couverture propres au pipeline sont générés lors de chaque analyse. Il y a quatre différents groupes d'indicateurs générés à différentes étapes du pipeline :

- Indicateurs d'alignement et de cartographie
- Indicateurs relatifs aux fichiers VCF
- Indicateurs de durée (ou de durée d'exécution)
- Rapports et indicateurs de couverture (ou d'enrichissement)

Les indicateurs d'alignement et de cartographie, les indicateurs relatifs aux fichiers VCF et un sous-ensemble des rapports sur la couverture disponibles sont automatiquement générés et ne nécessitent aucune activation ou commande particulière. Les autres indicateurs de couverture peuvent être activés. Des régions de couverture additionnelles peuvent également être spécifiées.

Les calculs des indicateurs sont effectués pendant l'analyse pour ne pas allonger la durée de l'analyse de la plateforme DRAGEN.

Figure 10 Génération d'indicateurs et de rapports



Format de sortie des indicateurs de CQ

Les indicateurs de CQ sont produits dans un format standard convivial et les fichiers CSV sont inscrits dans le répertoire de sortie de l'analyse.

- <output prefix>.mapping_metrics.csv
- <output prefix>.vc_metrics.csv
- <output prefix>.time_metrics.csv
- <output prefix>.<coverage region prefix>_coverage_metrics.csv
- <output prefix>.<other coverage reports>.csv

Chaque ligne est explicite, ce qui facilite la lecture.

Section	Région/ Échantillon	Indicateur	Dénombrement/ Rapport/Durée	Pourcentage/ Secondes
MAPPING/ALIGNING SUMMARY		Total input reads (Total de lectures en entrée)	816 360 354	
MAPPING/ALIGNING SUMMARY		Number of duplicate reads (marked not removed) (Nombre de lectures répétées [marquées mais non retirées])	15 779 031	1,93
...				
MAPPING/ALIGNING PER RG	RGID_1	Total reads in RG (Total de lectures de la région)	816 360 354	100
MAPPING/ALIGNING PER RG	RGID_1	Number of duplicate reads (marked) (Nombre de lectures répétées [marquées])	15 779 031	1,93
...				
VARIANT CALLER SUMMARY		Number of samples (Nombre d'échantillons)	1	
VARIANT CALLER SUMMARY		Reads Processed (Lectures traitées)	738 031 938	
...				
VARIANT CALLER PREFILTER	SAMPLE_ 1	Total	4 918 287	100

Section	Région/ Échantillon	Indicateur	Dénombrement/ Rapport/Durée	Pourcentage/ Secondes
VARIANT CALLER PREFILTER	SAMPLE_ 1	Biallelic (Biallélique)	4 856 654	98,75
...				
RUN TIME		Time loading reference (Durée de chargement de la référence)	00:18,6	18,65
RUN TIME		Time aligning reads (Durée d'alignement des lectures)	19:24,4	1164,42

Indicateurs d'alignement et de cartographie

Les indicateurs d'alignement et de cartographie, comme les indicateurs calculés par la commande Flagstat de Samtools, sont disponibles pour l'agrégat (toutes les données en entrée) et par groupe de lectures. Sauf indication contraire, les unités des indicateurs sont des lectures (et non des paires ou des alignements).

- Total input reads (Total de lectures en entrée) – Le nombre total de lectures dans les fichiers FASTQ en entrée.
- Number of duplicate marked reads (Nombre de lectures marquées comme répétition) – Les lectures marquées comme répétition quand l'option *--enable-duplicate-marking* est utilisée.
- Number of duplicate marked and mate reads removed (Nombre de lectures marquées comme répétitions et leur lecture couplée retirée) – Les lectures marquées comme répétition, avec l'autre lecture du couple, qui sont retirées quand l'option *--remove-duplicates* est utilisée.
- Number of unique reads (Nombre de lectures uniques) – Le nombre total de lectures moins les lectures marquées comme répétition.
- Reads with mate sequenced (Lectures couplées séquencées) – Le nombre de lectures avec une lecture couplée.
- Reads without mate sequenced (Lectures non couplées séquencées) – Le nombre total de lectures moins le nombre de lectures couplées séquencées.
- QC-failed reads (Lecture ayant échoué le CQ) – Les lectures qui n'ont pas passé les contrôles de qualité de la plateforme ou du fournisseur (marque SAM 0x200).
- Mapped reads (Lectures cartographiées) – Le nombre total de lectures cartographiées moins le nombre de lectures non cartographiées.
- Number of unique and mapped reads (Nombre de lectures cartographiées uniques) – Le nombre de lectures cartographiées moins le nombre de lectures marquées comme répétition.
- Unmapped reads (Lectures non cartographiées) – Le nombre total de lectures qui n'ont pu être cartographiées.
- Singleton reads (Lectures en singleton) – Le nombre de lectures cartographiées où l'autre lecture du couple n'a pu être lue.
- Paired reads (Lectures appariées) – Le dénombrement de lectures dont les deux lectures de la paire sont cartographiées.
- Properly paired reads (Lectures correctement appariées) – Les deux lectures de la paire sont cartographiées et se trouvent dans un intervalle acceptable l'une de l'autre selon l'estimation de la distribution de la longueur des inserts.
- Not properly paired reads (discordant) (Lectures incorrectement appariées [discordantes]) – Le nombre de lectures appariées moins le nombre de lectures correctement appariées.

- Paired reads mapped to different chromosomes (Lectures appariées cartographiées sur différents chromosomes) – Le nombre de lectures couplées dont l'autre lecture du couple est cartographiée sur un chromosome différent.
- Paired reads mapped to different chromosomes (MAPQ >= 10) (Lectures appariées cartographiées sur différents chromosomes [MAPQ >= 10]) – Le nombre de lectures couplées avec un score MAPQ supérieur à 10 dont l'autre lecture du couple est cartographiée sur un chromosome différent.
- Reads with indel R1 (Lectures R1 avec indel) – Le pourcentage des lectures R1 qui contiennent au moins 1 indel.
- Reads with indel R2 (Lectures R2 avec indel) – Le pourcentage des lectures R2 qui contiennent au moins 1 indel.
- Soft-clipped bases R1 (Bases R1 coupées conservées dans l'alignement de la lecture R1) – Le pourcentage de bases de la lecture R1 qui sont coupées, mais qui ont été conservées dans l'alignement.
- Soft-clipped bases R2 (Bases R2 coupées conservées dans l'alignement de la lecture R2) – Le pourcentage de bases de la lecture R2 qui sont coupées, mais qui ont été conservées dans l'alignement.
- Mismatched bases R1 (Bases mésappariées de la lecture R1) – Le nombre de bases mésappariées sur la lecture R1, qui est la somme du dénombrement de polymorphismes mononucléotidiques et des longueurs d'indels. Rien n'est dénombré dans les bases coupées, mais conservées dans l'alignement, ou dans les introns d'ARN. Le dénombrement n'est pas non plus effectué pour un mésappariement dont la base de référence ou de lecture est N.
- Mismatched bases R2 (Bases mésappariées de la lecture R2) – Le nombre de bases mésappariées sur la lecture R2, qui est la somme du dénombrement de polymorphismes mononucléotidiques et des longueurs d'indels. Rien n'est dénombré dans les bases coupées, mais conservées dans l'alignement, ou dans les introns d'ARN. Le dénombrement n'est pas non plus effectué pour un mésappariement dont la base de référence ou de lecture est N.
- Mismatched bases R1 (excluding indels) (Bases mésappariées de la lecture R1 [sans les indels]) – Le nombre de bases mésappariées sur la lecture R1. Les longueurs d'indels sont ignorées. Rien n'est dénombré dans les bases coupées, mais conservées dans l'alignement, ou dans les introns d'ARN. Le dénombrement n'est pas non plus effectué pour un mésappariement dont la base de référence ou de lecture est N.
- Mismatched bases R2 (excluding indels) (Bases mésappariées de la lecture R2 [sans les indels]) – Le nombre de bases mésappariées sur la lecture R2. Les longueurs d'indels sont ignorées. Rien n'est dénombré dans les bases coupées, mais conservées dans l'alignement, ou dans les introns d'ARN. Le dénombrement n'est pas non plus effectué pour un mésappariement dont la base de référence ou de lecture est N.
- Q30 Bases (Bases Q30) – Le nombre total de bases avec une valeur de BQ supérieure ou égale à 30.
- Q30 Bases R1 (Bases Q30 de lecture R1) – Le nombre total de bases de la lecture R1 avec une valeur de BQ supérieure ou égale à 30.
- Q30 Bases R2 (Bases Q30 de lecture R2) – Le nombre total de bases de la lecture R2 avec une valeur de BQ supérieure ou égale à 30.
- Q30 Bases (excluding dups & clipped bases) (Bases Q30 [sans les répétitions et les bases découpées]) – Le nombre de bases qui ne sont pas sur des répétitions et de bases non découpées avec une valeur de BQ supérieure ou égale à 30.
- Histogramme des qualités de la cartographie des lectures
  - Lectures avec score MAPQ [40:inf)
  - Lectures avec score MAPQ [30:40)

- Lectures avec score MAPQ [20:30)
- Lectures avec score MAPQ [10:20)
- Lectures avec score MAPQ [0:10)

- **Total alignments** (Total d'alignements) – Nombre total de lectures alignées à un locus avec une qualité supérieure à 0.
- **Secondary alignments** (Alignements secondaires) – Le nombre d'alignements secondaires d'un locus.
- **Supplementary (chimeric) alignments** (Alignements supplémentaires [chimériques]) – Une lecture chimérique fractionnée sur plusieurs locus (possiblement à cause de variants structurels). Un alignement est désigné comme alignement représentatif, les autres sont supplémentaires.
- **Estimated read length** (Estimation de la longueur des lectures) – Le nombre total de bases en entrée divisé par le nombre de lectures.
- **Histogram** (Histogramme) – Consultez la section *Rapport Histogram coverage (Histogramme de couverture)* à la page 109.
- **PCT of bases aligned that fell inside the interval region** (Pourcentage de bases alignées dans la région de l'intervalle) – Le nombre de bases dans la région de l'intervalle et la région cible divisé par le nombre total de bases alignées.
- **Estimated sample contamination** (Estimation de la contamination de l'échantillon) – Quand la définition de variants est effectuée avec un outil bioinformatique, la contamination entre les échantillons a une incidence importante sur l'exactitude de la prédiction. Même une très petite contamination peut entraîner la définition de nombreux faux positifs, particulièrement dans les pipelines visant à détecter des variants ayant une faible fréquence allélique.

  Le module de contamination entre les échantillons de la plateforme DRAGEN utilise un modèle de mélange probabiliste pour estimer la proportion des lectures d'un échantillon qui peut provenir d'une source humaine différente. Cette proportion de contamination d'échantillons est estimée comme une valeur de paramètre dans le modèle de mélange qui maximise la probabilité que les lectures soient observées à plusieurs emplacements d'empilement. Le modèle de mélange prend en compte les fréquences alléliques de la population et les génotypes d'échantillons inférés.

  Pour activer cet indicateur, vous devez fournir dans la ligne de commande le chemin d'accès vers un fichier VCF qui contient les sites de marqueurs (RSID) avec les fréquences alléliques de la population. Par exemple :

  ```
  --qc-cross-cont-vcf /opt/edico/config/sample_cross_contamination_
     resource_hg19.vcf
  ```

  Les fichiers VCF de ressources qui sont inclus avec la plateforme DRAGEN peuvent être reconstruits à partir de la base de données Ensembl. Les fichiers VCF inclus dans le répertoire de configurations de la plateforme DRAGEN contiennent environ 5000 emplacements de marqueurs dont la fréquence allélique est près de 0,5. Les fichiers sont propres à la référence (hg19/GRCh37/hg38). La plateforme DRAGEN interrompra le processus si des fichiers de ressources et de référence incompatibles sont utilisés (p. ex., le fichier de ressources CRCh37 avec le fichier de référence hg19).

  L'exemple suivant montre la sortie pour un échantillon sans contamination. La valeur est fournie sous forme de fraction, donc une valeur de « 0,011 » est équivalente à une estimation de contamination de 1,1 %.

  ```
  MAPPING/ALIGNING SUMMARY Estimated sample contamination 0.000
  ```

Indicateurs d'alignement et de cartographie du mode somatique

En mode somatique, les indicateur de cartographie et d'alignement sont générés séparément pour les échantillons de tumeur et de référence. Chaque ligne commence par TUMOR ou NORMAL, respectivement, pour indiquer de quel échantillon il s'agit. Les indicateurs de l'échantillon de tumeur sont d'abord produits, puis ceux de l'échantillon de référence le sont. Les indicateurs par groupe de lectures sont également séparés en groupes de lectures de tumeur et de référence.

Les indicateurs relatifs aux lectures alignées (couverture moyenne d'alignement du génome ou de la région cible, couverture moyenne de chromosome X ou Y du génome ou de la région cible, pourcentage de bases alignées dans l'intervalle de la région, etc.) ne sont PAS dénombrés séparément pour les échantillons de tumeur et de référence. C'est plutôt l'agrégat des deux échantillons qui est donné. Ces lignes ne commencent pas par TUMOR ou NORMAL et ne sont pas produites par groupe de lectures.

Indicateurs de définition de variants

Les indicateurs de définition de variants générés sont semblables aux indicateurs calculés par vcfstats de RTG. Les indicateurs sont fournis pour chaque échantillon dans les fichiers VCF et gVCF multiéchantillons. En fonction de l'analyse, les indicateurs sont fournis soit en format standard VARIANT CALLER ou JOINT CALLER. Les indicateurs sont fournis pour les fichiers VCF bruts (PREFILTER) et filtrés (POSTFILTER).

Les variants filtrés à l'aide d'un panel de référence et de la base de données COSMIC sont dénombrés comme des variants avec la marque « PASS » (réussite) dans les indicateurs POSTFILTER de fichier VCF. Ces variants portant la marque « PASS » (réussite) peuvent générer un dénombrement de variants plus élevé que prévu dans les indicateurs POSTFILTER de fichier VCF.

- Number of samples (Nombre d'échantillons) – Le nombre d'échantillons dans le fichier VCF combiné ou de population.
- Reads Processed (Lectures traitées) – Le nombre de lectures utilisées pour la définition de variants, en excluant toute lecture marquée comme répétition et les lectures qui se trouvent hors de la région cible.
- Total – Le nombre total de variants (polymorphismes mononucléotidiques + polymorphismes polynucléotidiques + indels).
- Biallelic (Biallélique) – Le nombre de sites d'un génome qui contient deux allèles observés, celui de référence étant dénombré comme un, ce qui permet un allèle de variant.
- Multiallelic (Multiallélique) – Le nombre de sites dans le fichier VCF qui contient trois allèles observés ou plus. Celui de référence est dénombré comme un, ce qui permet deux allèles de variant ou plus.
- SNPs (Polymorphismes mononucléotidiques) – Un variant est dénombré comme polymorphisme mononucléotidique quand les longueurs de la référence de l'allèle 1 et de l'allèle 2 sont toutes de 1.
- Insertions (Hom) – Le nombre de variants qui contient des insertions homozygotes.
- Insertions (Het) (Insertions [Hét]) – Le nombre de variants où les deux allèles sont des insertions, mais ne sont pas homozygotes.
- Deletions (Het) (Délétions [Hét]) – Le nombre de variants qui contient des délétions homozygotes.
- INDELS (Het) (Indels [Hét]) – Le nombre de variants dont le génotype est [insertion+délétion], [insertion+polymorphisme mononucléotidique] ou [délétion+polymorphisme mononucléotidique].
- De Novo SNPs (Polymorphismes mononucléotidiques de novo) – Les polymorphismes mononucléotidiques marqués *de novo* et ont une valeur de DQ supérieure à 0,05. Mettez la valeur de l'option *--qc-snp-denovo-quality-threshold* au seuil requis. La valeur par défaut est « 0,05 ».

- **De Novo INDELs** (Indels de novo) – Les indels marqués *de novo* avec une valeur de DQ supérieure à 0,02. Ce seuil de DQ peut être spécifié en mettant à l'option *--qc-indel-denovo-quality-threshold* la valeur de seuil de DQ requise. La valeur par défaut est « 0,02 ».
- **De Novo MNPs** (Polymorphismes polynucléotidiques de novo) – La même chose que l'option pour les polymorphismes mononucléotidiques. Définissez la valeur de l'option *--qc-snp-denovo-quality-threshold* au seuil requis. La valeur par défaut est « 0,05 ».
- **(Chr X SNPs)/(Chr Y SNPs) ratio in the genome (or the target region)** (Rapport entre les polymorphismes mononucléotidiques du chromosome X et ceux du chromosome Y dans le génome [ou la région cible]) – Le nombre de polymorphismes mononucléotidiques sur le chromosome X (ou à l'intersection du chromosome X et de la région cible) divisé par le nombre de polymorphismes mononucléotidiques sur le chromosome Y (ou à l'intersection du chromosome Y et de la région cible). S'il n'y a pas d'alignement au chromosome X ou Y, la valeur de cet indicateur est « NA ».
- **SNP Transitions** (Transitions de polymorphisme mononucléotidique) – Une permutation de deux purines (A<->G) ou de deux pyrimidines (C<->T).
- **SNP Transversions** (Transversions de polymorphisme mononucléotidique) – Une permutation de bases de purine et de pyrimidine. Rapport Ti/Tv : rapport des transitions aux transitions.
- **Heterozygous** (Hétérozygote) – Le nombre de variants hétérozygotes.
- **Homozygous** (Homozygote) – Le nombre de variants homozygotes.
- **Het/Hom ratio** (Rapport Hét/Hom) – Le rapport entre le nombre de variants hétérozygotes et homozygotes.
- **In dbSNP** (Dans dbSNP) – Le nombre de variants détectés qui sont présents dans le fichier de référence dbSNP. Si aucun fichier dbSNP n'a été fourni à l'aide de l'option *--bsnp*, la valeur des deux indicateurs **In dbSNP** et **Novel** est « NA ».
- **Novel** (Nouveaux) – Le nombre total de variants moins le nombre de variants dans dbSNP.
- **Percent Callability** (Pourcentage de définissabilité) – Disponible seulement dans le mode germinal avec la sortie de fichier gVCF. Le pourcentage de positions de référence non N qui ont « PASS » (réussite) comme valeur de typage génotypique. Les variants multialléliques ne sont pas dénombrés. Les délétions sont dénombrées pour toutes les positions de référence supprimées, mais seulement pour les définitions homozygotes. Seuls les autosomes et les chromosomes X, Y et M sont pris en compte.
- **Percent Autosome Callability** (Pourcentage de définissabilité des autosomes) – Seuls les autosomes sont pris en compte.
- **Percent QC Region Callability in Region i (i is equivalent to regions 1,2, or 3)** (Pourcentage de définissabilité de région de QC dans la région i [où i est l'équivalent des régions 1, 2 ou 3]) – Disponible si la définissabilité pour des régions personnalisées est demandée à l'aide de l'option *--qc-coverage-region-i* et que la sortie de la définissabilité est spécifiée grâce à l'option *--qc-coverage-reports-i*. Tous les contigs sont pris en compte.

Rapport sur la définissabilité

La plateforme DRAGEN génère automatiquement un rapport sur la définissabilité comme indicateur de la fonction de définition de variants lorsque la fonction de définition de petits variants germinaux est exécutée en mode de sortie gVCF. La définissabilité est définie comme étant la fraction des positions de référence non-N qui ont « PASS » (réussite) comme valeur de typage génotypique. Les critères suivants sont utilisés pour calculer les indicateurs de définissabilité :

- La définissabilité est calculée sur l'ensemble du fichier gVCF.
- Les contigs de leurres sont ignorés.

- Les contigs non positionnés et non localisés sont ignorés.
- Les positions N sont considérées comme étant non définissables.
- Pour les régions dans lesquelles aucune définition de variants n'a été effectuée, la définissabilité est nulle.
- Une délétion homozygote compte comme un typage génotypique « PASS » (réussite) pour toutes les positions de référence couvertes par la délétion.

Si l'option *--vc-target-bed* est spécifiée, le fichier de sortie porte l'extension target_bed_callability.bed et contient une définissabilité globale et autosomique dans la région cible du fichier BED d'entrée. La taille de l'élargissement spécifiée par l'option *--vc-target-bed-padding* est utilisée et les régions se chevauchant sont fusionnées.

La définissabilité peut aussi être inscrite dans les rapports personnalisés sur la couverture et la définissabilité.

Indicateurs de durée

La section sur les indicateurs de durée contient des données détaillées sur la durée d'exécution de chaque processus. Par exemple, les indicateurs suivants sont générés pour le pipeline de la fonction de cartographie et de définition de variants :

- Time loading reference (Durée de chargement de la référence)
- Time aligning reads (Durée d'alignement des lectures)
- Time sorting and marking duplicates (Durée du tri et du marquage des répétitions)
- Time DRAGStr calibration (Durée de la calibration de DRAGStr)
- Time partial reconfiguration (Durée de la reconfiguration partielle)
- Time variant calling (Durée de la définition des variants)
- Total runtime (Durée d'exécution totale)

Rapports sur la couverture et la définissabilité pour des régions personnalisées

La plateforme DRAGEN génère les rapports sur la couverture suivants :

- Un ensemble de rapports par défaut pour le génome entier ou, si l'option *--vc-target-bed* est spécifiée, pour la région cible.
- Des rapports supplémentaires (facultatifs) pour jusqu'à trois régions d'intérêt (régions de couverture).

  Pour chaque région spécifiée, la plateforme DRAGEN génère les rapports par défaut, ainsi que tout autre rapport facultatif demandé pour la région.

Pour générer des rapports sur les régions de couverture, utilisez l'option *–qc-coverage-region-i*, où i a une valeur de « 1 », « 2 » ou « 3 ».

- Chaque option *–qc-coverage-region-i* nécessite un fichier BED comme argument.
- Les régions de chaque fichier BED peuvent être élargies à l'aide de l'option *--qc-coverage-region-padding-i*, dont la valeur par défaut est « 0 ».
- Un ensemble de rapports par défaut est généré pour chaque région.
- Des rapports supplémentaires peuvent être spécifiés pour chaque région grâce à l'option *–qc-coverage-reports-i*.

L'exemple suivant montre les options requises pour générer des rapports sur la couverture.

```
$ dragen … \
   --qc-coverage-region-1 <bed file 1> \
   --qc-coverage-reports-1 full_res \
   --qc-coverage-region-2 <bed file 2> \
   --qc-coverage-region-3 <bed file 3> \
   --qc-coverage-reports-3 full_res cov_report
```

Dénombrement de lectures et de bases

Tous les rapports sur la couverture par défaut et facultatifs présentés dans le Tableau 6 et le Tableau 7 ci-dessous utilisent les règles suivantes par défaut pour dénombrer les lectures et les bases :

- Les lectures répétées sont ignorées.
- Les bases coupées présentes ou absentes dans l'alignement sont ignorées.
- Les lectures pour lesquelles le score de qualité MAPQ est égal à zéro sont ignorées.
- Les couples qui se chevauchent sont dénombrés deux fois.

Paramètres qui ne sont pas définis par défaut :

Les rapports sont disponibles avec ou sans l'exécution de la fonction de cartographie et d'alignement et la fonction de définition de variants. Toutefois, la valeur « true » (activée) (la valeur par défaut) doit être spécifiée pour les options *--enable-sort*.

Toute combinaison de rapports facultatifs peut être demandée pour chaque région. Si plusieurs types de rapports sont sélectionnés pour chaque région, ils devraient être séparés par des espaces.

Il est possible de remplacer les valeurs MAPQ (score de qualité) et BQ (qualité de base) minimales à appliquer pour une région donnée à l'aide de l'option qc-coverage-filters.

Un filtre de couverture est activé à l'aide des options *--qc-coverage-filters-i* (où i a une valeur de « 1 », « 2 » ou « 3 »), en combinaison avec l'option *--qc-coverage-region-i* correspondante :

- *--qc-coverage-region-i=<targetedregions.bed>*
- *--qc-coverage-filters-i <filters string>*

Par exemple, les options suivantes sont utilisées pour activer la production d'un fichier de couverture d'une résolution d'une paire de bases avec filtres :

```
--qc-coverage-region-1 <targetedregions.bed>
--qc-coverage-filters-1 'mapq<10,bq<30'
--qc-coverage-reports-1 full_res
```

- La syntaxe de l'argument est « mapq<value,bq<value », ce qui signifie que les lectures dont la qualité de la cartographie est inférieure à la valeur spécifiée ne sont pas dénombrées, de même que les bases dont la qualité de définition est inférieure à la valeur spécifiée.
- Les arguments de filtre valides sont « mapq » et « bq » seulement. Il est possible de spécifier l'un des deux ou les deux à la fois.
- Seul l'opérateur « < » est pris en charge. Les opérateurs « <= », « > », « >= » et « = » ne sont pas pris en charge.
- Lorsque le filtrage est activé pour une région cible, la plateforme DRAGEN produit les fichiers de rapports filtrés pour cette région. Les fichiers de rapports non filtrés ne sont pas produits pour la région filtrée ciblée.

Définissabilité pour les régions personnalisées

Si l'option *--qc-coverage-region-i* est utilisée avec *--qc-coverage-reports-i* (où i a une valeur de « 1 », « 2 » ou « 3 »), la définissabilité peut être ajoutée comme type de rapport pour cette région. Le fichier produit est un fichier qc-coverage-region-i_callability.bed. Pour chaque fichier qc-coverage-region-i spécifiée, la définissabilité moyenne est inscrite dans le fichier des indicateurs de définition de variants. La taille élargie spécifiée par l'option *--qc-coverage-region-padding-i* est utilisée et les régions se chevauchant sont fusionnées.

Les filtres de scores MAPQ et BQ minimaux n'ont un effet que sur le dénombrement de lectures et de bases. Ils n'ont aucun effet sur les rapports de définissabilité.

Les longueurs de contigs et de régions d'intérêt (utilisées comme dénominateurs) ne comprennent pas les régions dont le fichier FASTA comprend un « N ».

Types de rapports disponibles

Tableau 6 Rapport par défaut

Nom du rapport	Nom ou type de fichier de sortie
Coverage metrics (Indicateurs de couverture)	_coverage_metrics.csv
Fine histogram coverage (Histogramme de couverture détaillé)	_fine_hist.csv
Histogram coverage (Histogramme de couverture)	_hist.csv
Overall mean coverage (Couverture globale moyenne)	_overall_mean_cov.csv
Per contig mean coverage (Couverture moyenne par contig)	_contig_mean_cov.csv
Predicted ploidy (Prédiction de la ploïdie)	_ploidy.csv

Tableau 7 Rapports facultatifs

Nom du rapport	Type de fichier de sortie
full_res (résultats complets)	_full_res.bed
cov_report (couverture)	_cov_report.bed
callability (définissabilité)	_callability.bed

Rapport Coverage metrics (Indicateurs de couverture)

Le rapport sur les indicateurs de couverture produit un fichier _coverage_metrics.csv qui fournit des indicateurs concernant une région (la région peut être le génome, une région cible ou une région de couverture du contrôle de qualité (CQ). La première colonne du fichier de sortie contient le nom de section COVERAGE SUMMARY (résumé de la couverture). La deuxième colonne est réservée aux indicateurs.

Les critères suivants sont utilisés lors du calcul de la couverture :

- Les lectures répétées et les bases coupées sont ignorées.
- Seules les lectures dont le score MAPQ est supérieur au score MAPQ minimal et les bases dont le score BQ est supérieur au score BQ minimal sont prises en considération.

Les indicateurs suivants sont calculés :

- Aligned bases in region (Bases alignées dans la région) – Le nombre de bases à cartographie unique dans la région et pourcentage relatif au nombre de bases à cartographie unique par rapport au génome.
- Average alignment coverage over region (Couverture d'alignement moyenne de la région) – Le nombre de bases à cartographie unique dans la région divisé par le nombre de sites dans la région.
- Uniformity of coverage (PCT > 0.2*mean) over region (Uniformité de la couverture [pourcentage > 0,2*moyenne] de la région) – Le pourcentage des sites pour lesquels la couverture est plus de 20 % supérieure à la couverture moyenne dans la région.
- PCT of region with coverage [ix, inf) (Pourcentage de la région avec une couverture [ix, inf)) – Le pourcentage des sites de la région ayant une couverture d'au moins ix, où i peut valoir « 100 », « 50 », « 20 », « 15 », « 10 », « 3 », « 1 » ou « 0 ».
- PCT of region with coverage [ix, jx) (Pourcentage de la région avec une couverture [ix, jx)) – Le pourcentage des sites de la région ayant une couverture d'au moins jx, où (i, j) peut valoir « (50, 100) », « (20, 50) », « (15, 20) », « (10, 15) », « (3, 10) », « (1, 3) » ou « (0, 1) ».
- Average chromosome X coverage over region (Couverture moyenne de chromosome X de la région) – Le nombre total de bases alignées à l'intersection du chromosome X avec la région divisé par le nombre total de locus dans l'intersection du chromosome X avec la région. Si aucun chromosome X ne se trouve dans le génome de référence, ou que la région ne croise aucun chromosome X, cet indicateur affiche « NA » (s. o.).
- Average chromosome Y coverage over region (Couverture moyenne de chromosome Y de la région) – Le nombre total de bases alignées à l'intersection du chromosome Y avec la région divisé par le nombre total de locus dans l'intersection du chromosome Y avec la région. Si aucun chromosome Y ne se trouve dans le génome de référence, ou que la région ne croise aucun chromosome Y, cet indicateur affiche « NA » (s. o.).
- XAvgCov/YAvgCov ratio over genome/target region (Rapport de CouvMoyX/CouvMoyY de la région cible ou du génome) – La couverture moyenne de l'alignement du chromosome X dans la région divisée par la couverture moyenne de l'alignement du chromosome Y dans la région. Si aucun chromosome X ou Y ne se trouve dans le génome de référence, ou que la région ne croise aucun chromosome X ou Y, cet indicateur affiche « NA » (s. o.).
- Average mitochondrial coverage over region (Couverture mitochondriale moyenne de la région) – Le nombre total de bases alignées à l'intersection du chromosome mitochondrial avec la région divisé par le nombre total de locus dans l'intersection du chromosome mitochondrial avec la région. Si aucun chromosome mitochondrial ne se trouve dans le génome de référence, ou que la région ne croise aucun chromosome mitochondrial, cet indicateur affiche « NA » (s. o.).
- Average autosomal coverage over region (Couverture autosomale moyenne de la région) – Le nombre total de bases alignées avec le locus autosomique dans la région divisé par le nombre total de locus autosomiques dans la région. Si aucun locus autosomique ne se trouve dans le génome de référence, ou que la région ne croise aucun locus autosomique, cet indicateur affiche « NA » (s. o.).
- Median autosomal coverage over region (Couverture autosomale médiane de la région) – La couverture médiane de l'alignement par rapport au locus autosomique dans la région. Si aucun locus autosomique ne se trouve dans le génome de référence, ou que la région ne croise aucun locus autosomique, cet indicateur affiche « NA » (s. o.).
- Mean/Median autosomal coverage ratio over region (Rapport de couverture autosomale moyenne/médiane de la région) – La couverture moyenne du locus autosomique dans la région divisée

par la couverture médiane du locus autosomique dans la région. Si aucun locus autosomique ne se trouve dans le génome de référence, ou que la région ne croise aucun locus autosomique, cet indicateur affiche « NA » (s. o.).

- Aligned reads in region (Lectures alignées dans la région) – Le nombre de lectures à cartographie unique dans la région et pourcentage relatif au nombre de lectures à cartographie unique par rapport au génome. Lorsque la région correspond au fichier BED de cibles, cet indicateur est équivalent à la spécificité de capture (et la remplace) selon la région cible.

L'exemple suivant montre le contenu d'un fichier _coverage_metrics.csv :

```
COVERAGE SUMMARY,,Aligned bases,148169295474
COVERAGE SUMMARY,,Aligned bases in genome,148169295474,100.00
COVERAGE SUMMARY,,Average alignment coverage over genome,46.08
COVERAGE SUMMARY,,Uniformity of coverage (PCT > 0.2*mean) over genome,91.01
COVERAGE SUMMARY,,PCT of genome with coverage [100x: inf),0.25
COVERAGE SUMMARY,,PCT of genome with coverage [ 50x: inf),50.01
COVERAGE SUMMARY,,PCT of genome with coverage [ 20x: inf),89.46
COVERAGE SUMMARY,,PCT of genome with coverage [ 15x: inf),90.51
COVERAGE SUMMARY,,PCT of genome with coverage [ 10x: inf),91.01
COVERAGE SUMMARY,,PCT of genome with coverage [ 3x: inf),91.69
COVERAGE SUMMARY,,PCT of genome with coverage [ 1x: inf),92.10
COVERAGE SUMMARY,,PCT of genome with coverage [ 0x: inf),100.00
COVERAGE SUMMARY,,PCT of genome with coverage [ 50x:100x),49.76
COVERAGE SUMMARY,,PCT of genome with coverage [ 20x: 50x),39.45
COVERAGE SUMMARY,,PCT of genome with coverage [ 15x: 20x),1.04
COVERAGE SUMMARY,,PCT of genome with coverage [ 10x: 15x),0.51
COVERAGE SUMMARY,,PCT of genome with coverage [ 3x: 10x),0.67
COVERAGE SUMMARY,,PCT of genome with coverage [ 1x: 3x),0.42
COVERAGE SUMMARY,,PCT of genome with coverage [ 0x: 1x),7.90
COVERAGE SUMMARY,,Average chr X coverage over genome,24.70
COVERAGE SUMMARY,,Average chr Y coverage over genome,20.96
COVERAGE SUMMARY,,Average mitochondrial coverage over genome,20682.19
COVERAGE SUMMARY,,Average autosomal coverage over genome,47.81
COVERAGE SUMMARY,,Median autosomal coverage over genome,48.62
COVERAGE SUMMARY,,Mean/Median autosomal coverage ratio over genome,0.98
COVERAGE SUMMARY,,XAvgCov/YAvgCov ratio over genome,1.18
COVERAGE SUMMARY,,XAvgCov/AutosomalAvgCov ratio over genome,0.52
COVERAGE SUMMARY,,YAvgCov/AutosomalAvgCov ratio over genome,0.44
COVERAGE SUMMARY,,Aligned reads,1477121058
COVERAGE SUMMARY,,Aligned reads in genome,1477121058,100.00
```

Rapport Fine Histogram Coverage (Histogramme de couverture détaillé)

Le rapport Fine Histogram produit un fichier _fine_hist.csv contenant deux colonnes : Depth (profondeur) et Overall (global). La valeur indiquée dans la colonne Depth peut aller de « 0 » à « 1000+ » et la colonne Overall indique le nombre de locus couverts à la profondeur correspondante.

Rapport Histogram coverage (Histogramme de couverture)

Ce rapport d'histogramme produit un fichier _hist.csv qui fournit les suivants :

- Le pourcentage des bases dans la région du fichier BED de couverture, du fichier BED de cibles ou du séquençage de génome entier qui correspond à une certaine étendue de couverture.
- Les lectures répétées sont ignorées si l'option *--enable-duplicate-marking true* est utilisée.

Les étendues suivantes sont utilisées :

```
"[100x:inf)", "[1x:3x)", "[0x:1x)"
```

Rapport Overall mean coverage (Couverture globale moyenne)

Le rapport sur la couverture globale moyenne génère un fichier _overall_mean_cov.csv, qui contient la couverture moyenne de l'alignement du fichier BED de couverture, du fichier BED de cibles ou du séquençage de génome entier, le cas échéant.

L'exemple suivant montre le contenu d'un fichier _overall_mean_cov.csv :

```
Average alignment coverage over target_bed,80.69
```

Rapport sur la couverture moyenne par contig

Le rapport sur la couverture moyenne par contig génère un fichier _contig_mean_cov.csv qui contient l'estimation de la couverture pour l'ensemble des contigs ainsi que l'estimation de la couverture autosomique. Le fichier comprend les trois colonnes suivantes :

Colonne 1	Colonne 2	Colonne 3
Nom du contig	Le nombres de bases alignées avec ce contig, ce qui exclut les bases provenant de lectures marquées comme étant des répétitions, les lectures dont le score MAPQ est égale à zéro et les bases coupées.	L'estimation de la couverture, selon la formule suivante : <le nombre de bases alignées sur le contig (c.-à-d. Col2)> divisé par <la longueur du contig ou (si un fichier BED cible est utilisé) la longueur totale de la région cible couvrant le contig>.

Rapport Full Res (Résultats complets)

Le rapport des résultats complets est produit en format a _full_res.bed. Ce fichier est délimité par des tabulations et ses trois premières colonnes correspondent aux champs standards des fichiers BED. Sa quatrième colonne comprend la profondeur. Chaque enregistrement dans le fichier correspond à un intervalle donné à une profondeur constante. Si la profondeur change, un nouvel enregistrement est ajouté au fichier. Les alignements dont la valeur de la qualité de cartographie est « 0 », les lectures répétées et les bases coupées ne sont pas pris en compte dans le calcul de la profondeur.

Seules les positions de base qui appartiennent à la région de couverture spécifiée par l'utilisateur du fichier BED sont présentes dans le fichier de sortie _full_res.bed.

La structure du fichier _full_res.bed est semblable à celle du fichier de sortie *bedtools genomecov -bg*. Le contenu est identique si la ligne de commande « bedtools » est exécutée après avoir filtré les alignements dont le score de qualité de cartographie est « 0 » et possiblement après avoir filtré en utilisant un fichier BED de cibles (s'il est spécifié).

L'exemple suivant montre le contenu d'un fichier _full_res.bed :

```
chr1 121483984 121483985 10
chr1 121483985 121483986 9
chr1 121483986 121483989 8
```

```
chr1 121483989 121483991 7
chr1 121483991 121483992 6
chr1 121483992 121483993 4
chr1 121483993 121483994 2
chr1 121483994 121484039 1
chr1 121484039 121484043 2
chr1 121484043 121484048 3
chr1 121484048 121484050 7
chr1 121484050 121484051 11
chr1 121484051 121484052 17
chr1 121484052 121484053 149
chr1 121484053 121484054 323
chr1 121484054 121484055 2414
```

Rapport sur la couverture

Le rapport cov_report génère un fichier _cov_report.bed. Ce fichier est délimité par des tabulations et ses trois premières colonnes correspondent aux champs standards des fichiers BED. Les colonnes suivantes contiennent des statistiques calculées par rapport à la région de l'intervalle spécifiée sur la même ligne d'enregistrement. Les colonnes supplémentaires sont les suivantes :

- total_cvg – La valeur de couverture totale.
- mean_cvg – La valeur de couverture moyenne.
- Q1_cvg – La valeur de couverture du premier quartile (25e percentile).
- median_cvg – La valeur de couverture médiane.
- Q3_cvg – La valeur de couverture du dernier quartile (75e percentile).
- min_cvg – La valeur de couverture minimale.
- mean_cvg – La valeur de couverture maximale.
- pct_above_X – Indique le pourcentage des bases couvrant la région de l'intervalle spécifié ayant une couverture de profondeur supérieure à X.

Par défaut, si un intervalle a une couverture totale de « 0 », alors l'enregistrement est inscrit dans le fichier de sortie. Pour filtrer les intervalles dont la couverture est de « 0 », mettez à l'option *vc-emit-zero-coverage-intervals* la valeur « false » (désactivée) dans le fichier de configuration.

L'exemple suivant montre le contenu d'un fichier _cov_report.bed :

chrom	start	end	total_cvg	mean_cvg	Q1_cvg	median_cvg	Q3_cvg	min_cvg	max_cvg	pct_above_5
…										
chr5	34190121	34191570	76636	52.89	44.00	54.00	60.00	32	76	100.00
…										
chr5	34191751	34192380	39994	63.58	57.00	61.00	69.00	50	85	100.00
…										
chr5	34192440	34192642	10074	49.87	47.00	49.00	51.00	44	62	100.00
…										
chr9	66456991	66457682	31926	46.20	39.00	45.00	52.00	27	65	100.00
…										
chr9	68426500	68426601	4870	48.22	42.00	48.00	54.00	39	58	100.00
…										
chr17	41465818	41466180	24470	67.60	4.00	66.00	124.00	2	153	66.30
…										
chr20	29652081	29652203	5738	47.03	40.00	49.00	52.00	34	58	100.00
…										
chr21	9826182	9826283	4160	41.19	23.00	52.00	58.00	5	60	99.01
…										

Utilisation des rapports sur la couverture ou la définissabilité et fichiers de sortie attendus

Le tableau suivant montre quels fichiers de sortie sont générés lorsque les options par défaut (*--vc-target-bed*) ou les options facultatives relatives aux régions de couverture (*--coverage-region*) sont utilisées.

Option --vc-target-bed spécifiée? O/N	Option --qc- coverage-region-i (i a une valeur de 1, 2, ou 3) spécifiée? O/N	Fichiers de sortie attendus
N	N	wgs_coverage_metrics.csv wgs_fine_hist.csv wgs_hist.csv wgs_overall_mean_cov.csv wgs_contig_mean_cov.csv
N	O	wgs_coverage_metrics.csv wgs_fine_hist.csv wgs_hist.csv wgs_overall_mean_cov.csv wgs_contig_mean_cov.csv Pour chaque région de couverture spécifiée par l'utilisateur : qc-coverage-region-i_coverage_metrics.csv qc-coverage-region-i_fine_hist.csv qc-coverage-region-i_hist.csv qc-coverage-region-i_overall_mean_cov.csv qc-coverage-region-i_contig_mean_cov.csv qc-coverage-region-i_full_res.bed si le type de rapport full_res est demandé pour qc-coverage-region-i qc-coverage-region-i_cov_report.bed si le type de rapport cov_report est demandé pour qc-coverage-region-i qc-coverage-region-i_callability.bed si le mode GVCF est activé et que le type de rapport callability (définissabilité) ou exome-callability (définissabilité de l'exome) est demandé

Option --vc-target-bed spécifiée? O/N	Option --qc- coverage-region-i (i a une valeur de 1, 2, ou 3) spécifiée? O/N	Fichiers de sortie attendus
O	N	wgs_coverage_metrics.csv wgs_fine_hist.csv wgs_hist.csv wgs_overall_mean_cov.csv wgs_contig_mean_cov.csv target_bed_coverage_metrics.csv target_bed_fine_hist.csvtarget_bed_hist.csv target_bed_overall_mean_cov.csv target_bed_contig_mean_cov.csv target_bed_callability.bed si le mode GVCF est activé

Option --vc-target-bed spécifiée? O/N	Option --qc- coverage-region-i (i a une valeur de 1, 2, ou 3) spécifiée? O/N	Fichiers de sortie attendus
O	O	wgs_coverage_metrics.csv wgs_fine_hist.csv wgs_hist.csv wgs_overall_mean_cov.csv wgs_contig_mean_cov.csv target_bed_coverage_metrics.csv target_bed_fine_hist.csv target_bed_hist.csv target_bed_overall_mean_cov.csv target_bed_contig_mean_cov.csv target_bed_callability.bed si le mode GVCF est activé Pour chaque région de couverture spécifiée par l'utilisateur : qc-coverage-region-i _coverage_metrics.csv qc-coverage-region-i_fine_hist.csv qc-coverage-region-i_hist.csv qc-coverage-region-i_overall_mean_cov.csv qc-coverage-region-i_contig_mean_cov.csv qc-coverage-region-i_full_res.bed si le type de rapport full_res est demandé pour qc-coverage-region-i qc-coverage-region-i_cov_report.bed si le type de rapport cov_report est demandé pour qc-coverage-region-i qc-coverage-region-i_callability.bed si le mode GVCF est activé et que le type de rapport callability (définissabilité) ou exome-callability (définissabilité de l'exome) est demandé

Compression des indicateurs de couverture au format bigWig

Le fichier BED de sortie (_full_res.bed) contenant les indicateurs de couverture à résolution d'une paire de bases peut devenir très volumineux. La compression au format bigWig du fichier de sortie peut être activée en mettant à l'option *--enable-metrics-compression* la valeur « true » (activée).

Rapport de biais GC

Le rapport de biais GC fournit des renseignements sur la teneur en GC et sur la couverture de lecture associée dans un génome. L'indicateur de biais GC de la plateforme DRAGEN suit le modèle de l'outil Picard et est adapté aux mesures internes préexistantes.
Le module de correction des biais GC de la plateforme DRAGEN tente de corriger ces biais après l'étape du dénombrement de cibles. Pour plus de renseignements, consultez la section *Correction des biais GC* à la page 64.

L'indicateur de biais GC est calculé comme suit :

1 La teneur en GC est calculée à l'aide d'une fenêtre coulissante par base, sur 100 paires de bases, pour l'ensemble des chromosomes du génome de référence, excluant les contigs de leurres et les contigs ALT. Les fenêtres contenant plus de quatre bases masquées (N) dans la référence sont rejetées.

2 La couverture moyenne est calculée pour chaque fenêtre, excluant les lectures ne passant pas le filtre et les lectures répétées, secondaires ou supplémentaires.

3 La couverture moyenne globale à l'échelle du génome est calculée.

4 Les fenêtres valides selon la teneur (en pourcentage) en GC sont regroupées, à la fois sous forme de pourcentages individuels et sous forme de 5 fourchettes récapitulatives de 20 %.

5 La couverture normalisée pour chaque groupe est calculée en divisant la couverture moyenne du compartiment par la couverture moyenne globale à l'échelle du génome. Les valeurs inférieures à « 1,0 » indiquent une couverture inférieure à celle prévue dans le pourcentage ou la fourchette de GC donné. Les couvertures déviant de beaucoup de « 1,0 » et présentant des valeurs de GC élevées sont des résultats attendus.

6 Les indicateurs d'éliminations sont calculés comme la somme de toutes les valeurs positives (pourcentage de fenêtres à la position GC X - pourcentage de lectures alignées à la position GC X) chaque fois que la valeur de GC est inférieure ou égale à 50 % et pour chaque élimination AT et GC lorsque la valeur est supérieure à 50 %.

Par défaut, le rapport d'indicateurs de biais GC n'est pas calculé. Pour activer les calculs de biais GC, saisissez l'option de ligne de commande *--gc-metrics-enable* . Voici un exemple :

```
$ dragen -b <BAM file> -r <reference genome> --gc-metrics-enable=true
```

Le rapport d'indicateurs GC génère un fichier gc_metrics.csv. Ce fichier est structuré de la manière suivante :

```
GC BIAS DETAILS,,Windows at GC [0-100],<number of windows>,<fraction of
   all windows>
GC BIAS DETAILS,,Normalized coverage at GC [0-100],<average coverage of
   all windows at given GC divided by average coverage of whole genome>
GC METRICS SUMMARY,,Window size,<window size in base, typically 100>
GC METRICS SUMMARY,,Number of valid windows,<total number of windows used
   in calculations>
GC METRICS SUMMARY,,Number of discarded windows,<total number windows
   discarded due to more than 4 masked bases>
GC METRICS SUMMARY,,Average reference GC,<average GC content over all
   valid windows>
GC METRICS SUMMARY,,Mean global coverage,<average genome coverage over
   all valid windows>
```

```
GC METRICS SUMMARY,,Normalized coverage at GCs <GC range>,<average
   coverage of all windows at given GC range divided by average coverage
   of whole genome>
GC METRICS SUMMARY,,AT Dropout,<Calculated AT dropout value>
GC METRICS SUMMARY,,GC Dropout,<Calculated GC dropout value>
```

Le rapport de biais GC comprend aussi les options de lignes de commande suivantes. Il n'est toutefois pas recommandé de les utiliser.

- *--gc-metrics-window-size* – Remplace la fenêtre coulissante de 100 paires de bases par défaut.
- *--gc-metrics-num-bins* – Remplace le nombre de compartiments.

Rapport sur les indicateurs somatiques

En mode relatif aux échantillons de tumeur et de référence, la plateforme DRAGEN génère des rapports distincts pour l'échantillon de tumeur seulement et les échantillons de tumeur et de référence. Chaque rapport est étiqueté selon le type d'échantillon. La mention normal est ajoutée à la fin du nom de fichier s'il s'agit de rapports sur les échantillons de tumeur et de référence et la mention tumor (tumeur) est ajoutée à la fin du nom de fichier s'il s'agit d'un échantillon de tumeur seulement. Pour produire les résultats de l'échantillon de tumeur seulement et des échantillons de tumeur et de référence dans un seul fichier, mettez à l'option *--vc-enable-separate-t-n-metrics* la valeur « false » (désactivée). La plateforme DRAGEN produit alors des rapports sur l'agrégat des deux types d'échantillons.

La plateforme DRAGEN génère les rapports suivants :

- coverage_metrics
- gc_metrics
- contig_mean_cov
- fin_hist
- hist
- overall_mean_cov
- ploidy
- full_res (si demandé)
- ploidy (si demandé)

Si le fichier BED cible ou l'option *--qc-region-coverage-region-i* (où i est égal à 1, 2 ou 3) est inclus dans l'analyse, alors les rapports sur la couverture correspondants sont également divisés en fichiers pour l'échantillon de tumeur seulement et les échantillons de tumeur et de référence.

Rapport sur les régions somatiques définissables

En mode somatique, la plateforme DRAGEN génère automatiquement un fichier BED de régions somatiques définissables. Le rapport sur les régions somatiques définissables comprend les régions de l'échantillon dépassant les seuils de couverture spécifiés pour les échantillons de tumeur seulement et d'échantillons de tumeur et de référence. Chaque ligne du fichier BED de sortie est formatée comme suit :

```
chromosome region_start region_end
```

Vous pouvez spécifier la valeur du seuil à l'aide de l'option *--vc-callability-tumor-thresh* ou de l'option *--vc-callability-normal-thresh*. La valeur par défaut pour le seuil d'échantillon de tumeur seulement est de 15 et le seuil pour les échantillons de tumeur et de référence est de 5. Pour en savoir plus sur chaque option, consultez la section *Options du mode somatique* à la page 43.

Si le fichier BED cible ou l'option *--qc-region-coverage-region-i* (où i a une valeur de 1, 2 ou 3) est inclus dans l'analyse, la plateforme DRAGEN génère les fichiers BED des régions somatiques définissables correspondantes en plus du fichier BED des régions somatiques définissables du génome entier.

Détection virtuelle de lectures longues

La détection virtuelle de lectures longues (VLRD) de la plateforme DRAGEN est une autre fonction de définition de variants plus précise qui se concentre sur le traitement de régions homologues ou semblables du génome. Une fonction de définition de variants classique repose sur la fonction de cartographie et d'alignement pour déterminer quelles lectures proviennent d'un emplacement donné. Elle détecte également la séquence sous-jacente à cet emplacement indépendamment des autres régions qui n'y sont pas immédiatement adjacentes. La définition des variants classique fonctionne bien lorsque la région d'intérêt ne ressemble à aucune autre région du génome à l'échelle d'une seule lecture (ou d'une paire de lectures dans le cas du séquençage de lectures appariées).

Toutefois, une fraction importante du génome humain ne répond pas à ce critère. Plusieurs régions du génome ont des copies presque identiques ailleurs. C'est pourquoi la véritable source d'une lecture peut faire l'objet d'une incertitude considérable. Si un groupe de lectures est cartographié avec une confiance faible, une fonction de définition de variants typique pourrait l'ignorer, même s'ils contiennent des renseignements pertinents. Si une lecture est mal cartographiée (c.-à-d. si l'alignement primaire n'est pas la véritable source de la lecture), cela peut entraîner des erreurs de détection. Les technologies de séquençage de lectures courtes sont particulièrement susceptibles de faire face à ces problèmes. Le séquençage de lectures longues peut les atténuer, mais les coûts et les taux d'erreurs qui y sont associés sont généralement beaucoup plus élevés, ainsi que d'autres lacunes.

La fonction VLRD de la plateforme DRAGEN tente de s'attaquer aux complexités présentées par la redondance du génome à partir d'une perspective axée sur les données de lectures courtes. Plutôt que de prendre en considération chaque région de manière isolée, la fonction VLRD prend en considération tous les emplacements à partir desquels un groupe de lecture peut avoir été produit et tente de détecter les séquences sous-jacentes combinées à l'aide des renseignements disponibles.

Exécution de la fonction de détection de lectures longues virtuelles (VLRD) de la plateforme DRAGEN

Comme la fonction de définition de variants de la plateforme DRAGEN, la fonction VLRD produit un fichier VCF de sortie à partir d'un fichier BAM trié ou FASTQ d'entrée. La fonction VLRD prend en charge le traitement d'un ensemble de deux régions homologues seulement. La plateforme DRAGEN ne permet pas de traiter un ensemble de trois régions homologues ou plus. La prise en charge de trois régions homologues ou plus sera ajoutée dans une version ultérieure.

La fonction VLRD n'est pas activée par défaut. Pour l'exécuter, mettez à l'option *--enable-vlrd* la valeur « true » (activée). Voici un exemple de commande de la plateforme DRAGEN permettant d'exécuter la fonction VLRD :

```
dragen \
   -r <REF> \
   -1 <FQ1> \
    -2 <FQ2> \
   --RGID <RG> --RGSM <SM> \
   --output-dir <OUTPUT> \
   --output-file-prefix <PREFIX> \
   --enable-map-align true \
   --enable-sort=true \
```

```
--enable-duplicate-marking true \
--enable-vlrd true
--vc-target-bed similar_regions.bed
```

Fichier de sortie de cartographies et d'alignements mis à jour de la fonction VLRD

La fonction VLRD de la plateforme DRAGEN peut produire un fichier BAM ou SAM recartographié en plus du fichier de cartographie et d'alignement régulier. Pour activer la production d'un fichier BAM ou SAM recartographié, mettez à l'option *--enable-vlrd-map-align-output* la valeur « true » (activée). La valeur par défaut pour cette option est « false » (désactivée).

Le fichier de sortie supplémentaire de cartographies et d'alignement produit par la fonction VLRD ne contient que des lectures cartographiées dans les régions traitées par la fonction.

La fonction VLRD prend en considération les renseignements disponibles de toutes les régions homologues combinées pour mettre à jour les alignements des lectures (position ou qualité de la cartographie et ainsi de suite). Le fichier de cartographies et d'alignements mis à jour de la fonction VLRD est principalement utile à l'analyse d'empilements impliquant des régions homologues.

Paramètres de la fonction VLRD

Les options suivantes sont propres à la fonction VLRD dans le logiciel hôte de la plateforme DRAGEN :

- *--enable-vlrd*

  Lorsque la valeur est « true » (activée), la fonction VLRD est activée pour le pipeline de la plateforme DRAGEN.

- *--vc-target-bed*

  Spécifie le fichier BED d'entrée. La plateforme DRAGEN nécessite un fichier BED d'entrée spécifiant les régions homologues à traiter par la fonction VLRD. Ce fichier BED d'entrée doit avoir un format spécial requis par la fonction VLRD afin de bien traiter les régions homologues. L'étendue de régions maximale à traiter par la fonction VLRD est de 900 paires de bases.

  Par exemple :

  ```
  chr1    161497562   161498362   0   0
  chr1    161579204   161580004   0   0
  chr1    21750837     21751637   1   0
  chr1    21809355     21810155   1   1
  ```

  - Les trois premières colonnes correspondent à celles des fichiers BED classiques : la colonne 1 présente la description du chromosome, la colonne 2 le début de la région et la colonne 3 la fin de la région.
  - La colonne 4 contient l'identifiant de groupe des régions homologues. Il regroupe les régions qui sont homologues les unes aux autres.
  - Les lignes 1 et 2 ont la même valeur dans la colonne 4, ce qui indique qu'elles devraient être traitées comme un ensemble de régions homologues, indépendamment du groupe suivant (les lignes 3 et 4). Si les valeurs ne sont pas définies correctement, le logiciel pourrait regrouper des régions qui ne sont pas homologues les unes aux autres, ce qui amènerait des définitions de variants incorrectes.
  - La colonne 5 indique si une région est un complément inverse par rapport à l'autre région homologue. Une valeur de « 1 » dénote que la région est un complément inverse par rapport à d'autres régions du même groupe.

- À la ligne 4, la colonne 5 indique une valeur de « 1 ». Cela indique que la région est homologue à la région de la ligne 3 seulement s'il s'agit d'un complément inverse.

L'ensemble d'installation de la plateforme DRAGEN contient deux fichiers BED de la fonction VRLD pour les génomes de référence hg19 et hs37d5 sous /opt/edico/examples/VLRD. Vous pouvez utiliser ces fichiers BED tels quels pour exécuter la fonction VLRD ou comme modèle pour générer un fichier BED personnalisé.

- *--enable-vlrd-map-align-output*

  Lorsque la valeur est « true » (activée), la fonction VLRD produit un fichier BAM ou SAM cartographié qui ne contient que des lectures cartographiées dans les régions traitées par la fonction.

Génotypage forcé

La plateforme DRAGEN prend maintenant en charge le génotypage forcé (ForceGT) pour la définition de VNC en mode germinal. Pour utiliser ForceGT, définissez l'option *--vc-forcegt-vcf* avec une liste de petits variants pour forcer le génotypage. La liste d'entrée peut être au format .vcf ou .vcf.gz.

Les limites actuelles de ForceGT sont les suivantes :

- ForceGT est pris en charge pour la définition de VNC en mode germinal V3. Les modes V1, V2 et V2+ ne sont pas pris en charge.
- ForceGT n'est pas pris en charge pour la définition de VNC en mode somatique.
- Les variants produits par ForceGT ne sont pas générés par le génotypage combiné.

Fichier d'entrée de ForceGT

Le logiciel DRAGEN ne prend en charge qu'un seul fichier VCF d'entrée ForceGT qui doit satisfaire aux exigences suivantes :

- Contenir les mêmes contigs de référence que le fichier VCF utilisé pour la définition de variants.
- Être trié par nom et position de contigs de référence.
- Être normalisé (parcimonie et alignement à gauche).
- Ne contenir aucun variant complexe (des variants qui nécessitent plus d'une substitution/insertion/délétion pour passer de l'allèle REF à l'allèle ALT). Par exemple, n'importe quel variant du fichier VCT de ForceGT qui est semblable au variant suivant produira un comportement indéfini dans le logiciel de la plateforme DRAGEN :

  ```
  chrX 153592402 GTTGGGGATGCTGAC CACCCTGAAGGG
  ```

Les variants non normalisés suivants causent un comportement indéfini dans le logiciel DRAGEN :

- Sans parcimonie : `chrX 153592402 GC GCG`
- Représentation parcimonieuse : `chrX 153592403 C CG`

Fonctionnement de ForceGT et résultats attendus

En exécutant la définition de VNC en mode germinal avec ForceGT, un fichier gVCF uniéchantillon est généré à partir du fichier VCF d'entrée de ForceGT inscrit dans la ligne de commande DRAGEN. Le fichier gVCF uniéchantillon produit contient toutes les définitions régulières et les définitions ForceGT, comme suit :

- Pour une définition ForceGT non produite par la fonction de définition de variants (non commun), le champ INFO de la définition contient la mention « FGT ».

- Pour une définition ForceGT également produite par la fonction de définition de variants et pour laquelle le champ FILTER (filtre) indique « PASS » (réussite) (commun), le champ INFO de la définition présente l'étiquette « NML:FGT » (NML signifie « normal »).
- Pour une définition de référence (et « PASS ») (réussite) par la fonction de définition de variants, sans définition (de référence) de ForceGT, aucune étiquette n'est ajoutée (pas d'étiquette « NML » ni « FGT »).

Ce plan nous permet de faire la distinction entre les définitions présentes grâce à ForceGT seulement, commun aux fichiers d'entrée et aux définitions de référence de ForceGT et aux définitions de référence.

Tous les variants des fichiers VCF de ForceGT sont génotypés et présents dans le fichier gVCF uniéchantillon de sortie. Le génotypage rapporté pour les variants va comme suit :

Condition	Génotypage rapporté
À une position sans couverture	./.
À une position avec couverture, mais sans lecture en soutien à l'allèle ALT	0/0
À une position avec couverture et avec lecture en soutien à l'allèle ALT	0/0 ou 0/1 ou 1/1 ou 1/2

À une position où la définition de variants faite par défaut par le logiciel DRAGEN est différent de celle spécifiée dans le fichier VCF d'entrée de ForceGT, plusieurs entrées se trouvent à la même position dans le fichier gVCF de sortie, comme suit :

- Une entrée pour la définition de variants par défaut de la plateforme DRAGEN, et
- Une entrée pour chaque définition de variants présents dans le fichier VCF de ForceGT à cette position.

```
chrX 100 G C [définition de variants de la plateforme DRAGEN par
   défaut]
chrX 100 G A [variants dans le fichier VCF de ForceGT]
```

Si un fichier BED de cibles est fourni avec le fichier VCF de ForceGT, alors le fichier gVCF de sortie ne contient que les variants de ForceGT qui chevauchent les positions du fichier BED.

Identifiants moléculaires uniques

La plateforme DRAGEN peut traiter des données de génome entier et de tests de capture hybride à l'aide des identifiants moléculaires uniques (IMU). Les IMU sont des étiquettes moléculaires ajoutées aux fragments d'ADN avant l'amplification pour identifier la molécule d'ADN d'entrée originale des fragments amplifiés. Les IMU permettent de réduire les erreurs et les biais introduits par l'amplification ou le séquençage.

La plateforme DRAGEN prend en charge les types d'IMU suivants :

- IMU non aléatoires doubles comme ceux des réactifs IMU TruSight Oncology (TSO).
- IMU aléatoires de lectures à paire de bases unique comme les codes à barres moléculaires (MBC) SureSelect d'Agilent.

La plateforme DRAGEN utilise la séquence d'IMU pour regrouper les paires de lectures selon leur fragment d'entrée original et générer une paire de lectures de consensus pour chacun de ses groupes, ou familles. Le consensus permet de réduire les taux d'erreurs afin de détecter avec une grande exactitude des variants somatiques rares et à fréquence peu élevée dans les échantillons d'ADN. La plateforme DRAGEN génère un consensus comme suit :

1 Alignement des lectures.

2 Regroupement des lectures en groupes ayant des IMU et des alignements de paires identiques. Ces groupes sont appelés des familles.

3 Génération d'une seule paire de lectures de consensus pour chaque famille de lectures.

Ces lectures générées ont des scores de qualité plus élevés que les lectures d'entrée et reflètent une plus grande fiabilité, obtenue grâce à la combinaison de plusieurs observations de chaque définition de base.

La séquence d'IMU doit être présente dans le nom de la lecture. Pour inclure le nom de la lecture, spécifiez le paramètre OverrideCycles approprié dans l'outil de conversion en format BCL de la plateforme DRAGEN (consultez la section *Conversion des données en format BCL d'Illumina* à la page 154).

Pour utiliser le pipeline d'IMU, les fichiers de lectures en entrée doivent provenir d'une analyse à lectures appariées. La séquence d'IMU doit être dans le huitième des champs séparés par des deux-points de l'identifiant de la séquence. Les IMU peuvent avoir plusieurs parties qui sont séparées par le caractère « + ». Actuellement, la plateforme DRAGEN prend en charge les IMU ayant deux parties, chacune d'au plus 8 paires de bases, ou les IMU ayant une partie d'au plus 15 paires de bases. L'exemple qui suit montre l'IMU des deux lectures de la paire qui a été ajouté à la fin du nom de la lecture :

```
NDX550136:7:H2MTNBDXX:1:13302:3141:10799:AAGGATG+TCGGAGA
```

Options relatives aux IMU

- *--umi-enable* – Pour activer la combinaison de lectures, mettez à l'option *--umi-enable* la valeur « true » (activée). Cette option n'est pas compatible avec l'option *–enable-duplicate-marking*, car le pipeline d'IMU génère une lecture de consensus à partir d'un ensemble de lectures candidates en entrée plutôt que de choisir la meilleure lecture non répétée.
- *–umi-min-supporting-reads* – Utilisez l'option *–umi-min-supporting-reads* pour spécifier le nombre de lectures en entrée avec IMU identiques nécessaires pour générer une lecture de consensus. Toute famille qui n'a pas suffisamment de lectures à l'appui est rejetée. Sur la plateforme DRAGEN, il est recommandé de tester avec une plage de valeurs appropriée à votre préparation de librairies et profondeur de séquençage.

  Pour des renseignements sur l'évaluation des statistiques d'IMU observés, consultez la section *Indicateurs relatifs aux IMU* à la page 124.

Correction d'IMU aléatoires et non aléatoires

La plateforme DRAGEN traite les IMU est regroupant les lectures par IMU et par position d'alignement. S'il y a des erreurs dans les IMU, la plateforme DRAGEN peut détecter et corriger les petites erreurs de séquençage en utilisant une table de recherche ou la similarité entre les séquences et les dénombrements de lectures. Vous spécifiez le type de correction à l'aide de l'option *--umi-correction-scheme* en indiquant les valeurs « lookup » (recherche), « random » (aléatoire) ou « none » (aucun).

Pour les ensembles clairsemés d'IMU non aléatoires, il est possible de créer une table de recherche qui spécifie quelle séquence peut être corrigée ainsi que comment la corriger. Ce programme de fichier de correction fonctionne le mieux sur des ensembles d'IMU où les séquences ont une petite distance de Hamming ou d'édition entre elles. Par défaut, la plateforme DRAGEN utilise la correction par recherche avec une table de correction intégrée pour la solution TruSight Oncology d'Illumina et les trousses IDT pour ancres d'index d'IMU d'Illumina. Spécifiez le chemin d'accès de votre fichier de correction grâce à l'option *--umi-correction-file*. Si vous utilisez un ensemble différent d'IMU non aléatoires, communiquez avec l'assistance technique d'Illumina pour obtenir des renseignements sur la génération du fichier de correction correspondant.

Dans le programme de correction d'IMU aléatoires, la plateforme DRAGEN doit inférer quels IMU d'une position donnée sont probablement des erreurs par rapport aux autres IMU trouvés à la même position. La plateforme DRAGEN accomplit ceci comme suit.

- Les lectures sont regroupées par position d'alignement de fragment.
- Au sein d'une fenêtre floue à chaque position, les lectures sont regroupées par séquence d'IMU, ce qui constitue une famille.
- Si deux IMU se trouvent a une distance de Hamming de 1 l'un de l'autre et qu'une famille a moins de lectures que l'autre, les lectures de la plus petite famille sont soit fusionnées à la famille plus grande soit supprimées. La plus petite famille est fusionnée quand il n'y a pas d'autre grande famille à une distance de Hamming de 1 afin qu'elle puisse être corrigée sans ambiguïté. Vous pouvez spécifier la différence minimale de coefficient entre le nombre de paires de lectures des familles pour que la fusion soit effectuée grâce à l'option *--umi-random-merge-factor*.

Fusion d'IMU de duplex

Les adaptateurs d'IMU de duplex étiquettent simultanément les deux brins de fragments d'ADN double brin. Il est alors possible d'identifier les lectures provenant d'une amplification de chaque brin du fragment initial.

La plateforme DRAGEN considère que deux paires de lectures combinées sont la séquence des deux brins du même fragment d'ADN initial s'ils ont la même position d'alignement (au sein d'une fenêtre floue), s'ils ont des orientations complémentaires et si leurs IMU ont été échangés de la lecture 1 à la lecture 2. Si vous activez la fusion de duplex avec l'option *--umi-enable-duplex-merging*, alors les lectures d'entrée de ces deux groupes correspondants sont combinées et la plateforme DRAGEN générera une seule lecture de consensus.

Exemple de commande d'identifiants moléculaires uniques (IMU)

L'exemple suivant montre une ligne de commande de la plateforme DRAGEN permettant de générer un fichier BAM à partir des lectures d'entrée en consensus avec des IMU d'Illumina :

```
dragen \
   -r <REF> \
   -1 <FQ1> \
   -2 <FQ2> \
   --output-dir <OUTPUT> \
   --output-file-prefix <PREFIX> \
   --enable-map-align true \
   --enable-sort true \
   --umi-enable true \
   --umi-correction-scheme=lookup \
   --umi-min-supporting-reads 2
```

Pour lancer une analyse à partir d'IMU aléatoires, mettez à l'option *--umi-correction-scheme* la valeur « random » (aléatoire). Vous pouvez aussi utilisez l'option *--umi-random-merge-factor* pour spécifier le rapport du nombre de fragments nécessaire pour corriger et fusionner deux familles ayant des IMU semblables.

Indicateurs relatifs aux IMU

La plateforme DRAGEN produit un fichier <output_prefix>.umi_metrics.csv qui fournit les statistiques pour la combinaison des IMU. Ce fichier résume les statistiques sur les lectures d'entrée, comment elles ont été regroupées en famille, comment les IMU ont été corrigés et comment les familles ont généré des lectures de consensus. Les indicateurs suivants peuvent être utiles quand vous effectuez un réglage fin du pipeline pour votre application :

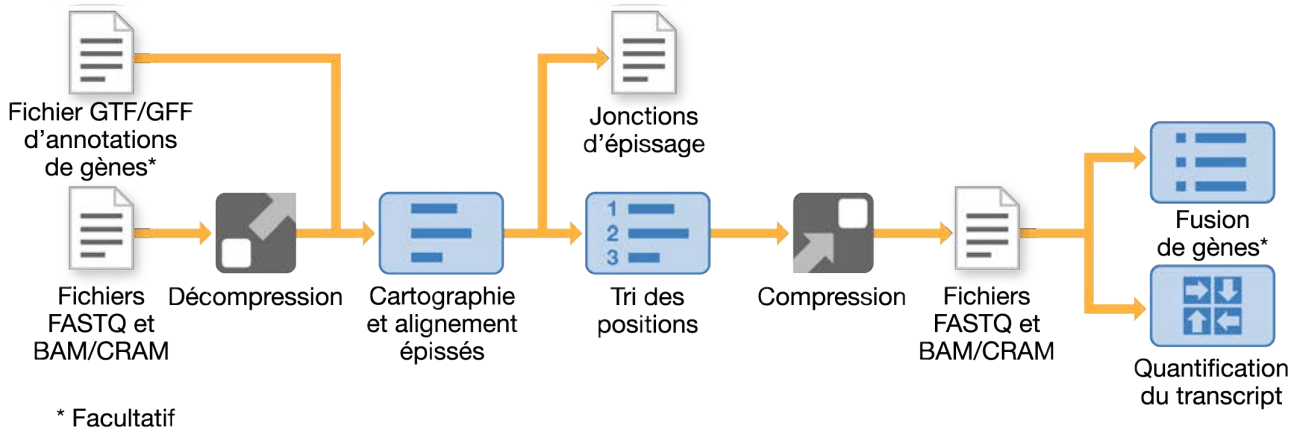
- Discarded families (Familles rejetées) – Toute famille qui a moins de fragments d'entrée que la valeur de l'option *--umi-min-supporting-reads* est rejetée. Ces lectures sont enregistrées comme Reads filtered out (Lectures filtrées). Les familles sont enregistrées comme Families discarded (Familles rejetées).
- UMI correction (Correction d'IMU) – Les familles peuvent être combinées de différentes façons. Le nombre de ces corrections est signalé comme suit :
  - Families shifted (Familles déplacées) – Les familles se trouvant à une distance égale ou moindre à celle spécifiée par le paramètre *umi-fuzzy-window-size*. La valeur par défaut du paramètre *umi-fuzzy-window-size* est « 3 ».
  - Families contextually corrected (Familles avec correction contextuelle) – Les familles avec exactement les mêmes coordonnées d'alignement de fragment et des IMU compatibles sont fusionnées.
  - Duplex families (Familles de duplex) – Les familles avec des coordonnées d'alignement similaires et des IMU complémentaires sont fusionnées.

Lorsque vous spécifiez un chemin d'accès valide pour l'option *--umi-metrics-interval-file*, la plateforme DRAGEN produit un ensemble séparé de statistiques d'IMU exactes qui ne contient que les familles qui se trouvent dans le fichier BED spécifié.

Si vous devez analyser l'étendue de la couverture par les IMU observés de l'ensemble de l'espace des séquences d'IMU possibles, l'indicateur Histogram of unique UMIs per fragment position (Histogramme des IMU uniques par position de fragment) peut être utile. Il s'agit d'un histogramme à coordonnées commençant à zéro où l'index indique un dénombrement d'IMU uniques à une position de fragment spécifique. Sa valeur représente le nombre de positions avec ce dénombrement.

Chapitre 4 Pipeline d'ARN de la plateforme DRAGEN

La plateforme DRAGEN comprend une fonction d'alignement de séquençage de l'ARN (sensible à l'épissage), ainsi que des composants d'analyse propres à l'ARN pour la quantification d'expression génique et la détection de fusions de gènes.



La plupart des fonctionnalités et des options décrites dans les sections d'options du logiciel hôte et de cartographie de l'ADN s'appliquent également aux applications utilisant de l'ARN. Les aspects additionnels propres à l'ARN sont décrits dans cette section.

Fichiers d'entrée

Fichier d'annotations de gènes

En plus des fichiers standards d'entrée (lectures d'un fichier fastq ou bam, génome de référence, etc.), la plateforme DRAGEN peut prendre en entrée un fichier d'annotations de gènes. Un fichier d'annotations de gènes aide à aligner les lectures à des jonctions d'épissage et sert à la quantification des expressions et à la détection de fusions de gènes.

Pour spécifier un fichier d'annotations de gènes, utilisez l'option de ligne de commande *-a* (*--annotation-file*). Le fichier d'entrée doit être conforme à la spécification du format GTF/GFF (http://uswest.ensembl.org/info/website/upload/gff.html). Le fichier doit contenir les caractéristiques du type d'exon et l'enregistrement doit contenir les attributs de type gene_id et transcript_id. Voici un exemple de fichier GTF valide :

```
chr1    HAVANA  transcript  11869   14409   .   +   .   gene_id
"ENSG00000223972.4"; transcript_id "ENST00000456328.2"; …
chr1    HAVANA  exon        11869   12227   .   +   .   gene_id
"ENSG00000223972.4"; transcript_id "ENST00000456328.2"; …
chr1    HAVANA  exon        12613   12721   .   +   .   gene_id
"ENSG00000223972.4"; transcript_id "ENST00000456328.2"; …
chr1    HAVANA  exon        13221   14409   .   +   .   gene_id
"ENSG00000223972.4"; transcript_id "ENST00000456328.2"; …
chr1    ENSEMBL transcript  11872   14412   .   +   .   gene_id
"ENSG00000223972.4"; transcript_id "ENST00000515242.2"; …
chr1    ENSEMBL exon        11872   12227   .   +   .   gene_id
```

```
"ENSG00000223972.4"; transcript_id "ENST00000515242.2"; …
chr1    ENSEMBL exon        12613   12721   .   +   .   gene_id
"ENSG00000223972.4"; transcript_id "ENST00000515242.2"; …
chr1    ENSEMBL exon        13225   14412   .   +   .   gene_id
"ENSG00000223972.4"; transcript_id "ENST00000515242.2"; …
…
```

Un fichier GFF peut également être utilisé. Chaque caractéristique d'exon doit avoir comme parent un identifiant de transcrit qui sert à regrouper les exons. Voici un exemple de fichier GFF valide :

```
1   ensembl_havana  processed_transcript    11869   14409       .   +   .
ID=transcript:ENST00000456328;
1   havana          exon                    11869   12227       .   +   .
Parent=transcript:ENST00000456328; …
1   havana          exon                    12613   12721       .   +   .
Parent=transcript:ENST00000456328; …
1   havana          exon                    13221   14409       .   +   .
Parent=transcript:ENST00000456328; …
…
```

Le logiciel hôte de la plateforme DRAGEN analyse le fichier pour trouver des exons au sein des transcrits et produire des jonctions d'épissage. La sortie suivante affiche le nombre de jonctions d'épissage qui ont été détectées.

```
================================================================
Generating annotated splice junctions
================================================================
Input annotations file: ./gencode.v19.annotation.gtf
Splice junctions database file: output/rna.sjdb.annotations.out.tab

 Number of genes: 27459

 Number of transcripts: 196520
 Number of exons: 1196293
 Number of splice junctions: 343856
```

Les jonctions d'épissage qui ont été détectées sont également inscrites dans un fichier de sortie, *.sjdb.annotations.out.tab, que l'utilisateur peut consulter. Lors de la formation de jonctions d'épissage à partir du fichier d'annotations de gènes en entrée, un simple filtre sert à éliminer les jonctions d'épissage qui n'ont pas la longueur minimale requise. Cela permet de réduire le taux de fausses détections de jonctions erronément annotées. La longueur minimale d'annotation de jonction d'épissage est spécifiée à l'aide de l'option *--rna-ann-sj-min-len*. Sa valeur par défaut est « 6 ».

Mode à deux passages

Plutôt que d'utiliser un fichier GTF/GFF pour les jonctions d'épissage annotées, le logiciel de la plateforme DRAGEN peut lire un fichier SJ.out.tab (consultez la section *Fichier SJ.out.tab* à la page 128). Ce fichier permet à la plateforme DRAGEN d'exécuter le mode à deux passages où les jonctions d'épissage découvertes lors du premier passage (produit en fichier SJ.out.tab) sont utilisées pour guider la cartographie

et l'alignement des lectures lors d'un second passage dans la plateforme DRAGEN. Ce mode de fonctionnement est utile pour augmenter la sensibilité envers les alignements épissés quand un fichier d'annotations de gènes n'est pas facile à obtenir pour le génome cible.

Alignement de l'ARN

Le pipeline d'ARN de la plateforme DRAGEN utilise la fonction d'alignement épissé de séquençage de l'ARN de la plateforme DRAGEN. La cartographie de séquences de point d'ancrage court à partir de lectures de séquençage de l'ARN est similaire à la cartographie de lectures d'ADN. Les jonctions d'épissage (la jonction des exons non contigus dans les transcrits d'ARN) à proximité des points d'ancrage cartographiés sont détectées et incorporées à l'ensemble des alignements de lectures.

Fichiers d'alignements de sortie

Les fichiers de sortie générés lors de l'exécution de la plateforme DRAGEN en mode ARN sont similaires à ceux générés en mode ADN. Le mode ARN produit également des renseignements additionnels à propos des alignements épissés. Les détails relatifs aux jonctions d'épissage sont présents dans l'enregistrement d'alignement du fichier SAM et dans un fichier additionnel, le fichier SJ.out.tab.

Fichier BAM

Le fichier BAM de sortie est conforme à la spécification SAM et est compatible avec les outils d'analyse de séquençage d'ARN en aval.

Étiquettes BAM de séquençage de l'ARN

Les étiquettes BAM suivantes sont produites avec les alignements épissés.

- XS:A – L'étiquette XS indique l'orientation du brin d'un intron. Consultez la section *Compatibilité avec Cufflinks* à la page 128.
- jM:B – L'étiquette jM indique les motifs d'intron pour toutes les jonctions des alignements. Elle a les définitions suivantes :
  - 0 : non canonique
  - 1 : GT/AG
  - 2 : CT/AC
  - 3 : GC/AG
  - 4 : CT/GC
  - 5 : AT/AC
  - 6 : GT/AT

  Si un fichier d'annotations de gènes est utilisé pendant l'étape de cartographie et d'alignement et que la jonction d'épissage est détectée comme jonction annotée, alors 20 est ajouté à la valeur du motif.

NH:i – Une étiquette SAM standard indiquant le nombre d'alignements signalés qui contiennent la requête dans l'enregistrement actuel. Cette étiquette peut être utilisée pour des outils en aval comme featureCounts.

HI:i – Une étiquette SAM standard dénotant l'index d'occurrences de la requête don la valeur indique que l'alignement est le *ie* stocké dans le fichier SAM. L'intervalle de sa valeur est de 1 à NH. Cette étiquette peut être utilisée pour des outils en aval comme featureCounts.

Compatibilité avec Cufflinks

Cufflinks pourrait avoir besoin d'alignements épissés pour produire l'étiquette de brin XS:A. Cette étiquette est présente dans l'enregistrement SAM si l'alignement contient une jonction d'épissage. Les valeurs de l'étiquette de brin XS:A sont les suivantes :

« . » (non défini), « + » (brin direct), « - » (brin inverse) ou « * » (ambigu).

Si l'alignement épissé a un brin non défini ou ambigu, alors l'alignement peut être supprimé en mettant à l'option *--no-ambig-strand* la valeur « 1 ».

Cufflinks s'attend également à ce que le champ MAPQ pour une lecture à cartographie unique ne contienne qu'une seule valeur. Cette valeur est spécifiée à l'aide de l'option *--rna-mapq-unique*. Pour forcer toutes les lectures à cartographie unique a ce que la valeur du champ MAPQ soit égale à cette valeur, mettez à l'option *--rna-mapq-unique* une valeur autre que zéro.

Fichier SJ.out.tab

En plus des alignements produits dans le fichier SAM ou BAM, un fichier SJ.out.tab additionnel résume les jonctions d'épissage à fiabilité élevée dans un fichier à valeurs séparées par des tabulateurs. Les colonnes de ce fichier sont les suivantes :

1 Le nom du contig

2 La première base de la jonction d'épissage (coordonnées commençant par 1)

3 La dernière base de la jonction d'épissage (coordonnées commençant par 1)

4 Brin (0 : non défini, 1 : +, 2 : -)

5 Le motif d'intron : 0 : non canonique, 1 : GT/AG, 2 : CT/AC, 3 : GC/AG, 4 : CT/GC, 5 : AT/AC, 6 : GT/AT

6 0 : non annoté, 1 : annoté, seulement si un fichier d'annotations de gènes est utilisé en entrée

7 Le nombre de lectures à cartographie unique englobant la jonction d'épissage

8 Le nombre de lectures à cartographies multiples englobant la jonction d'épissage

9 La longueur maximale d'extrémité saillante d'alignement épissé

Le champ de longueur maximale d'extrémité saillante d'alignement épissé (colonne 8) dans le fichier SJ.out.tab est l'extrémité saillante d'alignement d'ancrage. Par exemple, si une lecture est épissée comme ACGTACGT----------ACGT, alors l'extrémité saillante a une longueur de 4. Pour la même jonction d'épissage, sur toutes les lectures où se trouve cette jonction, la longueur maximale d'extrémité saillante est signalée. La longueur maximale d'extrémité saillante est un indicateur de fiabilité que la jonction d'épissage est correcte basée sur les alignements d'ancrage.

Deux fichiers SJ.out.tab sont générés par le logiciel hôte de la plateforme DRAGEN, une version non filtrée et une version filtrée. Les enregistrements du fichier non filtré sont une consolidation de tous les enregistrements d'alignements épissés du fichier SAM ou BAM de sortie. La version filtrée a toutefois une fiabilité beaucoup plus élevée quant à l'exactitude en raison de l'utilisation des filtres suivants.

Une entrée de jonction d'épissage dans le fichier SJ.out.tab ne passe pas le fichier si *une* de ces conditions est satisfaite :

- La jonction d'épissage a un motif non canonique et est justifiée par moins de 3 cartographies uniques.
- La jonction d'épissage a une longueur supérieure à 50 000 et est justifiée par moins de 2 cartographies uniques.

- La jonction d'épissage a une longueur supérieure à 100 000 et est justifiée par moins de 3 cartographies uniques.
- La jonction d'épissage a une longueur supérieure à 200 000 et est justifiée par moins de 4 cartographies uniques.
- La jonction d'épissage a un motif non canonique et la longueur maximale d'extrémité saillante d'alignement épissé est inférieure à 30.
- La jonction d'épissage a un motif canonique et la longueur maximale d'extrémité saillante d'alignement épissé est inférieure à 12.

L'utilisation du fichier SJ.out.tab filtré est recommandée pour toute analyse en aval ou outil post-traitement. Vous pouvez également utiliser le fichier SJ.out.tab non filtré et appliquer vos propres filtres (par exemple, avec les commande awk de base).

Veuillez prendre note que le filtre ne s'applique pas aux alignements se trouvant dans le fichier BAM ou SAM.

Fichier Chimeric.out.junction

S'il y a des alignements chimériques dans l'échantillon, alors un fichier Chimeric.out.junction supplémentaire est également produit. Ce fichier contient des renseignements à propos des lectures fractionnées qui peuvent servir pour effectuer la détection de fusions de gènes en aval. Chaque ligne contient une lecture d'alignement chimérique. Les colonnes du fichier sont les suivantes :

1 Le chromosome du donneur d'épissage.

2 La première base de l'intron du donneur d'épissage (coordonnées commençant par 1).

3 Le brin du donneur d'épissage.

4 Le chromosome de l'accepteur d'épissage.

5 La première base de l'intron de l'accepteur d'épissage (coordonnées commençant par 1).

6 Le brin de l'accepteur d'épissage.

7 S. O. – La colonne n'est pas utilisée, mais elle est présente pour la compatibilité avec d'autres outils. Sa valeur sera toujours « 1 ».

8 S. O. – La colonne n'est pas utilisée, mais elle est présente pour la compatibilité avec d'autres outils. Sa valeur sera toujours « * ».

9 S. O. – La colonne n'est pas utilisée, mais elle est présente pour la compatibilité avec d'autres outils. Sa valeur sera toujours « * ».

10 Le nom de la lecture.

11 La première base du premier segment, sur le brin +.

12 La chaîne CIGAR du premier segment.

13 La première base du deuxième segment.

14 La chaîne CIGAR du deuxième segment.

Les chaînes CIGAR de ce fichier sont conformes aux opérations standards de chaînes CIGAR telles que décrites dans la spécification SAM, avec un écart de longueur L ajouté qui est encodé avec l'opération p. Pour les lectures appariées, la séquence de la seconde lecture est toujours mise en ordre de complément inverse avant l'identification du brin.

L'exemple suivant montre un fichier d'entrée indiquant deux lectures appariées avec alignement chimérique, dont une des lectures est fractionnée, où les segments sont cartographiés sur les chromosomes 19 et 12. Les enregistrements SAM correspondants associés à ces entrées sont également indiqués.

```
chr19 580462 + chr12 120876182 + 1 * * R_15448 571532 49M8799N26M8p49M26S
120876183 49H26M
chr19 580462 + chr12 120876182 + 1 * * R_15459 571552 29M8799N46M8p29M46S
120876183 29H46M

R_15448:1   99    chr19   571531      60  49M8799N26M  =       580413
R_15448:2   147   chr19   580413      60  49M26S       =       571531
R_15448:2   2193  chr12   120876182   15  49H26M       chr19   571531

R_15459:1   99    chr19   571551      60  29M8799N46M  =       580433
R_15459:2   147   chr19   580433      4   29M46S       =       571551
R_15459:2   2193  chr12   120876182   15  29H46M       chr19   571551
```

Options d'alignement de l'ARN

L'étape d'alignement de la fonction d'alignement épissé de l'ARN utilise les options de notation d'alignement Smith-Waterman et les options de notation d'épissage.

Options relatives au score d'alignement Smith-Waterman

Consultez la section *Paramètres du score d'alignement Smith-Waterman* à la page 19 pour en savoir plus sur l'algorithme d'alignement utilisé par la plateforme DRAGEN. Les options suivantes relatives à la notation sont propres au traitement de motifs canoniques et non canoniques au sein des introns.

- *--Aligner.intron-motif12-pen*

  L'option *--Aligner.intron-motif12-pen* spécifie la pénalité pour les motifs canoniques 1/2 (GT/AG, CT/AC). La valeur par défaut calculée par le logiciel hôte est 1*(match-score + mismatch-pen).

- *--Aligner.intron-motif34-pen*

  L'option *--Aligner.intron-motif34-pen* spécifie la pénalité pour les motifs canoniques 3/4 (GC/AG, CT/GC). La valeur par défaut calculée par le logiciel hôte est 3*(match-score + mismatch-pen).

- *--Aligner.intron-motif56-pen*

  L'option *--Aligner.intron-motif56-pen* spécifie la pénalité pour les motifs canoniques 5/6 (AT/AC, GT/AT). La valeur par défaut calculée par le logiciel hôte est 4*(match-score + mismatch-pen).

- *--Aligner.intron-motif0-pen*

  L'option *--Aligner.intron-motif0-pen* spécifie la pénalité pour les motifs non canoniques. La valeur par défaut calculée par le logiciel hôte est 6*(match-score + mismatch-pen).

Options relatives à la notation d'épissage

- *--Mapper.min-intron-bases*

  Pour la cartographie de séquençage de l'ARN, un écart dans les alignements de la référence peut être interprété comme une délétion ou un intron. S'il n'y a pas de jonction d'épissage annotée, la valeur de l'option min-intron-bases constitue la longueur d'écart de seuil qui établit cette distinction. Des écarts de référence dont la longueur est supérieure à ce seuil sont interprétés et notés en tant qu'introns et les écarts de la référence plus courts sont interprétés et notés en tant que délétions. Toutefois, des alignements peuvent être retournés avec des jonctions d'épissage annotées plus courtes que ce seuil.

- *--Mapper.max-intron-bases*

  L'option max-intron-bases permet de décider la longueur maximale des introns qui sont signalés, ce qui est utile pour prévenir le signalement de fausses jonctions d'épissage. Mettez à cette option une valeur appropriée à l'espèce pour laquelle vous effectuez la cartographie.

- *--Mapper.ann-sj-max-indel*

  Pour le séquençage de l'ARN, la cartographie de points d'ancrage peut déceler un écart de la référence à la position d'un intron annoté lorsque sa longueur est légèrement différente. Si la différence de longueur n'est pas plus grande que la valeur de cette option, la fonction de cartographie vérifiera s'il est possible que l'intron soit présent exactement comme il est annoté, mais qu'un indel près de la jonction d'épissage, d'un côté ou de l'autre, puisse expliquer la différence de longueur. Il est très probable que les indels plus longs que la valeur de cette option et très près de la jonction d'épissage annotée ne soient pas détectés. Des valeurs plus élevées pourraient faire augmenter la durée de la cartographie et le nombre de fausses détections.

Notation du score MAPQ

Par défaut, le calcul du score MAPQ est identique pour le séquençage de l'ARN et le séquençage de l'ADN. Le contributeur principal au calcul du score MAPQ est la différence entre le meilleur score d'alignement et le deuxième meilleur score. Donc, modifier les paramètres de notation d'alignement a une incidence sur l'estimation du score MAPQ. Ces modifications sont expliquées dans la section *Paramètres du score d'alignement Smith-Waterman* à la page 19.

L'option *--mapq-strict-sjs* est propre à l'ARN et s'applique quand au moins un segment d'exon est aligné de manière fiable, mais qu'il y a une ambiguïté relative à des jonctions d'épissage possibles. Si la valeur de cette option est « 0 », un score MAPQ plus élevé est retourné, ce qui exprime avec confiance que l'alignement est au moins partiellement correct. Si la valeur de cette option est « 1 », un score MAPQ plus bas est retourné, ce qui exprime une ambiguïté de jonction d'épissage.

Certains outils en aval, comme Cufflinks, s'attendent à ce que la valeur de MAPQ soit unique pour toutes les lectures à cartographie unique. Cette valeur est spécifiée à l'aide de l'option *--rna-mapq-unique*. Mettre à cette option une valeur autre que zéro écrase toutes les estimations du score MAPQ basées sur le score d'alignement. Toutes les lectures à cartographie unique ont un score MAPQ équivalent à la valeur de l'option *--rna-mapq-unique*. Toutes les lectures à cartographies multiples ont une valeur de score MAPQ de int(-10*log10(1 - 1/NH), où la valeur de NH est le nombre d'occurrences (alignements primaires et secondaires) pour cette lecture.

Détection de fusions de gènes

Le module de fusions de gènes de la plateforme DRAGEN utilise la fonction d'alignement épissé de l'ARN de la plateforme DRAGEN pour détecter les événements de fusion de gènes. Il effectue une analyse des lectures fractionnées sur les alignements (chimériques) supplémentaires pour déceler les points de rupture potentiels. Les événements de fusion putative passent alors à travers différentes étapes de filtrage pour réduire le nombre de faux positifs potentiels. En plus des résultats définitifs, tous les candidats potentiels (non filtrés) sont produits, ce qui peut servir pour maximiser la sensibilité.

Exécution du module de fusions de gènes de la plateforme DRAGEN

Le module de fusions de gènes peut être exécuté conjointement à la tâche normale de cartographie et d'alignement de séquençage de l'ARN. Ce module additionnel n'ajoute que peu de traitement à la durée d'exécution globale, tout en fournissant des renseignements pour vos expériences de séquençage de l'ARN.

Pour activer le module de fusions de gènes de la plateforme DRAGEN, mettez à l'option *--enable-rna-quantification* la valeur « true » (activée) dans vos scripts de lignes de commande de séquençage de l'ARN actuels. Le module de fusions de gènes requiert l'utilisation du fichier d'annotations de gènes en format GTF ou GFF.

Ci-dessous se trouve un exemple de ligne de commande pour exécuter une expérience de séquençage de l'ARN de bout en bout.

```
/opt/edico/bin/dragen \
  -r <HASHTABLE> \
  -1 <FASTQ1> \
  -2 <FASTQ2> \
  -a <GTF_FILE> \
  --output-dir <OUT_DIRECTORY> \
  --output-file-prefix <PREFIX> \
  --RGID <READ_GROUP_ID> \
  --RGSM <Sample_NAME> \
  --enable-rna true \
  --enable-rna-gene-fusion true
```

À la fin d'une analyse, un résumé des événements de fusion de gènes semblable à l'exemple suivant est produit :

```
================================================================
Loading gene annotations file
================================================================
  Input annotations file: ref_annot.gtf
  Number of genes: 27459
  Number of transcripts: 196520
  Number of exons: 1196293

================================================================
Launching DRAGEN Gene Fusion Detection
================================================================
annotation-file:            ref_annot.gtf
rna-gf-blast-pairs:         blast_pairs.outfmt6
rna-gf-exon-snap:           50
rna-gf-min-anchor:          25
rna-gf-min-neighbor-dist:   15
rna-gf-max-partners:        3
rna-gf-min-score-ratio:     0.15
rna-gf-min-support:         2
rna-gf-min-support-be:      10
rna-gf-restrict-genes       true

================================================================
Completed DRAGEN Gene Fusion Detection
================================================================
Chimeric alignments: 107923
Total fusion candidates: 38 (2116 before filters)

Time loading annotations:                              00:00:08.543
Time running gene fusion:                              00:00:18.470
```

```
Total runtime:                                          00:00:27.760
************************************************************
DRAGEN finished normally
```

Exécution autonome du module de fusions de gènes

Le module de fusions de gènes de la plateforme DRAGEN peut être exécuté comme utilitaire autonome en prenant en entrée le fichier *.Chimeric.out.junction et le fichier d'annotations de gènes en format GTF ou GFF. Exécuter le module de fusions de gènes en utilitaire autonome est utile pour essayer différentes options de configuration à l'étape de détection de fusion de gènes sans avoir à cartographier et aligner plusieurs fois les données de séquençage de l'ARN.

Pour exécuter le module de fusions de gènes de la plateforme DRAGEN en utilitaire autonome, utilisez l'option *--rna-gf-input-file* pour spécifier le fichier *.Chimeric.out.junction déjà généré.

L'exemple montre une ligne de commande pour l'exécution du module de fusions de gènes en utilitaire autonome :

```
/opt/edico/bin/dragen \
   -a <GTF_FILE> \
   --rna-gf-input-file <INPUT_CHIMERIC> \
   --output-dir <OUT_DIRECTORY \
   --output-file-prefix <PREFIX> \
   --enable-rna true \
   --enable-rna-gene-fusion true
```

Le mode autonome ne produit pas les mêmes résultats que l'exécution à partir de lectures.

Candidats de fusion de gènes

Les fichiers de sortie fusion_candidate contiennent les événements de fusion de gènes détectés.
Le fichier de sortie contient les colonnes suivantes : Toute colonne additionnelle décrit les caractéristiques additionnelles des candidats de fusion.

- **#FusionGene** – Les noms des gènes parents (en ordre de transcrit de 5' à 3') qui participent à la fusion.
- **Score** – Le score de fusion basé sur les paires de lectures uniques qui supportent la fusion. Pour calculer le score, le nombre de paires de lectures fractionnées est ajouté au nombre de paires de lectures discordantes, puis divisé par 2.
- **LeftBreakpoint** – Le point de rupture du gène 1 dans le format <Chromosome>:<Position>:<Brin>.
- **RightBreakpoint** – Le point de rupture du gène 2 dans le format <Chromosome>:<Position>:<Brin>.
- **Filter** – La valeur `PASS` signifie que le candidat passe tous les critères de filtre (réussite). Si le candidat ne passe pas tous les critères de filtre, alors la valeur est « `FAIL;` » suivie d'une liste séparée par des points-virgules des critères utilisés pour filtrer le candidat.

Les candidats de fusion qui ont passé les étapes de filtrage sont énumérés en premier, en ordre décroissant de score.

Options et filtres de fusion de gènes

Différents filtres sont mis en œuvre pour réduire le nombre de candidats de fusion de gènes qui sont des faux positifs. Les options et les seuils suivants sont configurables. Ces filtres sont appliqués et tous les candidats qui se qualifient sont produits dans le fichier de sortie **.fusion_candidates.final*. Pour accéder à tous les candidats de fusion avant le filtrage, consultez le fichier de sortie **.fusion_candidates.preliminary*.

- *--rna-gf-blast-pairs*

  Un fichier contenant les paires de gènes qui ont une grande similarité. Cette liste de paires de gènes sert comme filtre d'homologie pour réduire le nombre de faux positifs. Une méthode pour générer ce fichier est de suivre les instructions décrites dans le Wiki de FusionFiltre (en anglais seulement). Utilisez le fichier `ref annot.cdsplus.fa.allvsall.outfmt6.genesym.gz` produit par la trousse d'outils CTAT. Dans les cas d'analyses de génome humain, un fichier par défaut est inclus et utilisé automatiquement si aucun autre ficher n'est spécifié manuellement.

- *--rna-gf-restrict-genes*

  Lors de l'analyse du fichier d'annotations de gènes (format GTF/GFF) à utiliser avec le module de fusions de gènes de la plateforme DRAGEN, cette option peut être utilisée pour restreindre les entrées d'intérêt aux régions de codage de protéine seulement. Restreindre le fichier GTF aux gènes de codage de protéine et d'ARNlinc afin de réduire les taux de faux positifs dans les événements de fusion actuellement à l'étude. La valeur par défaut est « true » (activée).

Quantification de l'expression génique

Le pipeline d'ARN de la plateforme DRAGEN contient un module de quantification de l'expression génique qui permet d'estimer l'expression de chaque transcrit et de chaque gène dans un ensemble de données de séquençage de l'ARN. Le module traduit d'abord la cartographie génomique de chaque lecture (ou paire de lectures) en cartographies de transcrit correspondantes. Il utilise ensuite un algorithme d'Espérance-Maximisation (EM) pour inférer les valeurs d'expression du transcrit qui correspondent le mieux avec les lectures observées. L'algorithme EM peut aussi modéliser le biais GC et apporter les corrections nécessaires dans le rapport des résultats de quantification.

Exécution de la quantification

Pour activer le module de quantification, mettez à l'option *--enable-rna-quantification* la valeur « true » (activée) dans les scripts actuels de vos lignes de commande de séquençage de l'ARN. De plus, la quantification nécessite l'utilisation d'un fichier d'annotation génique (GTF/GFF) qui fournit la position génomique de tous les transcrits à quantifier. Spécifiez-le dans l'option -a (ou *--annotation-file*).

Résultats de la quantification

Les résultats de la quantification de transcrits sont rapportés dans le fichier <outputPrefix>.quant.sf, un fichier texte contenant les résultats pour chaque transcrit. Par exemple :

```
Name                Length  EffectiveLength     TPM             NumReads
ENST00000364415.1   116     12.3238             5.2328          1
ENST00000564138.1   2775    2105.58             1.28293         41.8885
```

- Name – Indique l'identifiant du transcrit.
- Length – Indique la longueur du transcrit (épissé) en paires de bases.
- EffectiveLength – Indique la longueur accessible au séquençage de l'ARN, en prenant en compte la taille de l'insert et des effets de bord.
- TPM (transcrits par million) – Représente l'expression du transcrit, normalisée par rapport à la longueur du transcrit et à la profondeur du séquençage.
- NumReads – Indique l'estimation du nombre de lectures dans le transcrit (non normalisée).

Ce fichier peut être utilisé comme fichier d'entrée pour définir l'expression génique différentielle à l'aide d'outils comme tximport et DESeq2.

De la même manière, le fichier <outputPrefix>.quant.genes.sf contient des résultats de quantification au niveau du gène. Ces résultats sont produits en calculant tous les transcrits ayant le même identifiant (geneID) dans l'annotation (GTF). Les valeurs de « Length » et « EffectiveLength » représentent un moyen de pondérer l'expression du transcrit individuel dans le gène.

Options de quantification

- *--enable-rna-quantification*

  Si la valeur est « true » (activée), la quantification de l'ARN est activée. L'option *--enable-rna* doit également être spécifiée avec la valeur « true » (activée).

- *--rna-quantification-library-type*

  Spécifie le type de librairie de séquençage de l'ARN.
  - IU – La librairie appariée sans brin.
  - ISR – La librairie appariée à brins dans laquelle « read2 » correspond au brin du transcrit (p. ex., ARN TruSeq).
  - ISF – La librairie appariée à brins dans laquelle « read1 » correspond au brin du transcrit (p. ex., ARN TruSeq).
  - U – La librairie sans brin à paire de bases unique.
  - SR – La librairie à brins et à paire de bases unique dans laquelle les lectures suivent l'orientation inverse au brin du transcrit (p. ex., ARN TruSeq).
  - SF – La librairie à brins et à paire de bases unique dans laquelle les lectures correspondent au brin du transcrit.
  - A (autodétection, par défaut) – Pour cette valeur, la plateforme DRAGEN examine les premières lectures ou paires de l'ensemble de données pour déterminer automatiquement le bon type de librairie.

- *--rna-quantification-gc-bias*

  La correction de biais GC estime l'incidence du %GC du transcrit sur la couverture du séquençage et la prend en compte dans son estimation de l'expression. Mettre à cette option la valeur « false » (désactivée) permet de désactiver la correction de biais GC.

- *--rna-quantification-fld-max*, *--rna-quantification-fld-mean*, *--rna-quantification-fld-sd*

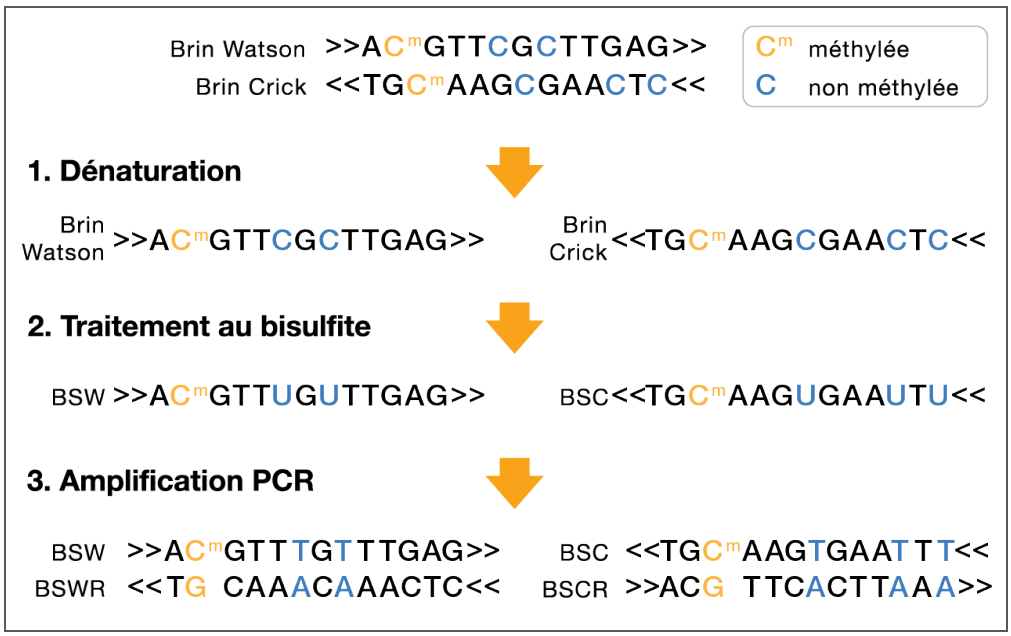
  Ces options servent à spécifier la distribution de la taille des inserts de la librairie de séquençage de l'ARN pour les analyses à paire de bases uniques, ce qui est pertinent pour la correction de biais GC.
  Les valeurs par défaut sont « 250 +- 25, max =1000 »; remplacez-les par des valeurs correspondant à la librairie en question pour en améliorer l'exactitude.

Chapitre 5 Pipeline de méthylation de la plateforme DRAGEN

La méthylation épigénétique des bases de cytosine dans l'ADN peut avoir un effet important sur l'expression génique. Le séquençage au bisulfite est la norme de référence pour la détection de modèles de méthylation épigénétique à une résolution d'une seule base. Cette technique utilise le traitement chimique de l'ADN avec du bisulfite de sodium, ce qui convertit les bases de cytosine non méthylées en uracile sans modifier les cytosines méthylées. L'amplification par PCR subséquente convertit les uraciles en thymines.

Une librairie de séquençage au bisulfite peut être directionnelle ou non directionnelle. Quand elle est non directionnelle, chaque fragment d'ADN double brin produit quatre brins distincts pour le séquençage après l'amplification, comme illustré dans la figure ci-dessous :

Figure 11 Séquençage au bisulfite non directionnel



- Brin Watson traité au bisulfite (BSW), complément inverse du brin BSW (BSWR),
- Brin Crick traité au bisulfite (BSC), complément inverse du brin BSC (BSCR)

Pour les librairies directionnelles, les quatre types de brins sont générés, mais des adaptateurs sont liés aux fragments d'ADN pour faire en sorte que seuls les brins BSW et BSC sont séquencés (protocole de Lister). Plus rarement, les brins BSWR et BSCR sont sélectionnés pour le séquençage (protocole directionnel avec compléments).

Brins BSW et BSC :

- A, G et T : inchangées
- La C méthylée reste de la C
- La C non méthylée est convertie en T

Brins BSWR et BSCR :

- Les bases complémentaires aux bases A, G et T des brins Watson et Crick initiaux demeurent inchangées
- La G complémentaire à la C méthylée des brins Watson et Crick initiaux reste en G

- La G complémentaire à la C non méthylée des brins Watson et Crick initiaux devient une A

Le séquençage d'ADN standard sert à produire les lectures de séquençage. Les lectures contenant plus de C non converties ou de G complémentaires aux C non converties sont beaucoup moins touchées par le traitement au bisulfite et ont une probabilité plus élevée d'être cartographiées avec la référence que les lectures dont plus de bases sont touchées. Le protocole standard pour minimiser ce biais de mise en correspondance est d'effectuer plusieurs alignements par lecture où des combinaisons spécifiques de lectures et de bases du génome de référence sont converties *in silico* avant chaque analyse d'alignement. Chaque analyse d'alignement a un ensemble de contraintes et de conversions de bases qui correspondent à un des types de brins traités au bisulfite et par PCR qui est attendu du protocole. En comparant les résultats des alignements de toutes les analyses, vous pouvez déterminer quel est le meilleur alignement et probablement le type de brins de chaque lecture ou paire de lectures. Ces renseignements sont nécessaires pour la définition de la méthylation en aval.

Définition de la méthylation de la plateforme DRAGEN

Les différents protocoles de méthylation requièrent la génération de deux ou quatre alignements par lecture en entrée, puis une analyse pour sélectionner le meilleur alignement et déterminer quelles cytosines sont méthylées. La plateforme DRAGEN peut automatiser ce processus en générant un seul fichier BAM de sortie présentant des étiquettes compatibles avec Bismark (XR, XG et XM) qui peut être utilisé dans des pipelines en aval comme les scripts d'extraction de la méthylation de Bismark.

Quand l'option *--methylation-protocol* a une valeur valide et non vide, la plateforme DRAGEN produit automatiquement l'ensemble d'alignements requis, chacun avec ses conversions appropriées sur les lectures, ses conversions sur la référence et ses contraintes quant à l'alignement dans l'ordre normal ou du complément inverse (CI) avec la référence. Quand l'option *--enable-methylation-calling* a la valeur « true » (activée), la plateforme DRAGEN analyse les différents alignements pour produire un seul fichier BAM avec des étiquettes de méthylation. Quand l'option *--enable-methylation-calling* a la valeur « false » (désactivée), la plateforme DRAGEN produit en sortie un fichier BAM pour chaque analyse d'alignement.

Le tableau suivant décrit ces analyses d'alignement :

Protocole	Fichier BAM	Référence	Lecture 1	Lecture 2	Contrainte d'orientation
Directionnel					
	1	C->T	C->T	G->A	ordre normal seulement
	2	G->A	C->T	G->A	ordre de CI seulement
Directionnel avec compléments ou non directionnel					
	1	C->T	C->T	G->A	ordre normal seulement
	2	G->A	C->T	G->A	ordre de CI seulement
	3	C->T	G->A	C->T	ordre de CI seulement
	4	G->A	G->A	C->T	ordre normal seulement

Dans les protocoles directionnels, la librairie est préparée de manière à ce que seuls les brins BSW et BSC soient séquencés. Les analyses d'alignement sont alors effectuées avec les deux combinaisons de conversions de bases et de contraintes d'orientation qui correspondent à ces brins (les analyses directionnelles 1 et 2 ci-dessus). Toutefois, avec les protocoles non directionnels, des lectures de chacun des quatre brins sont également probables. Des analyses d'alignement doivent donc être effectuées avec deux autres combinaisons de conversions de bases et de contraintes d'orientation (analyses non directionnelles 3 et 4 ci-dessus).

Le protocole **directionnel avec compléments** est semblable au protocole directionnel, sauf que les lectures de séquençage ne sont dérivées que des brins de compléments inverses des brins BSW et BSC. Dans le cas du protocole directionnel avec compléments, très peu de bons alignements sont attendus des lectures 1 et 2, la plateforme DRAGEN est donc automatiquement finement réglée pour utiliser un mode d'analyse plus rapide sur ces lectures.

Pour tous les protocoles, vous devez effectuer l'analyse en utilisant une référence générée avec l'option *--ht-methylated* activée comme décrit dans la section *Tables de hachage propres à un pipeline* à la page 153.

Voici un exemple de ligne de commande dragen pour le protocole directionnel :

```
dragen --enable-methylation-calling true \
   --methylation-protocol directional \
   --ref-dir /staging/ref/mm10/methylation --RGID RG1 --RGCN CN1 \
   --RGLB LIB1 --RGPL illumina --RGPU 1 --RGSM Samp1 \
   --intermediate-results-dir /staging/tmp \
   -1 /staging/reads/samp1_1.fastq.gz \
   -2 /staging/reads/samp1_2.fastq.gz \
   --output-directory /staging/outdir \
   --output-file-prefix samp1_directional_prot
```

Étiquettes de fichier BAM liées à la méthylation

Quand l'option *--enable-methylation-calling* a la valeur « true » (activée), la plateforme DRAGEN analyse les alignements produits pour le protocole configuré par l'option *--methylation-protocol* et génère un seul fichier BAM de sortie qui comprend les étiquettes liées à la méthylation pour toutes les lectures cartographiées.Comme avec Bismark, les lectures qui n'ont pas un meilleur alignement unique sont exclues du fichier BAM de sortie. Les étiquettes ajoutées sont les suivantes :

Étiquette	Description brève	Description
XR:Z	Conversion de lecture	Pour le meilleur alignement, la conversion de bases qui a été effectuée sur la lecture : CT ou GA.
XG:Z	Conversion de référence	Pour le meilleur alignement, la conversion de bases qui a été effectuée sur la référence : CT ou GA.
XM:Z	Définition de la méthylation	Une chaîne de méthylation d'un octet par base.

L'étiquette XM:Z (définition de la méthylation) contient un octet correspondant à chaque base dans la séquence de la lecture. Il y a un point (« . ») pour chaque position où il n'y a pas de cytosine et une lettre là où il y en a. La lettre indique le contexte (CpG, CHG, CHH ou inconnu). La casse précise s'il y a de la méthylation, c'est-à-dire qu'une majuscule indique une position méthylée et une minuscule une position non méthylée. Les lettres utilisées aux positions de la cytosine sont les suivantes :

Caractère	Méthylée?	Contexte
.	Pas de la cytosine	Pas de la cytosine
z	Non	CpG
Z	Oui	CpG
X	Non	CHG
X	Oui	CHG

Caractère	Méthylée?	Contexte
h	Non	CHH
H	Oui	CHH
u	Non	Inconnu
U	Oui	Inconnu

Rapports sur la méthylation de la cytosine et le biais M

Pour générer un rapport sur la méthylation de la cytosine pour l'ensemble du génome, mettez à l'option *--methylation-generate-cytosine-report* la valeur « true » (activée). La position et le brin de chaque C du génome sont indiqués dans les trois premiers champs du rapport. Un enregistrement présentant un « - » dans le champ du brin correspond à un G dans le fichier FASTA de référence. Les dénombrements de C méthylés et non méthylés qui couvrent la position sont fournis dans le quatrième et le cinquième champs, respectivement. Le contexte de C de la référence (CG, CHG ou CHH) est indiqué dans le sixième champ. Le contexte de la séquence trinucléotidique se trouve dans le dernier champ (p. ex., CCC, CGT, CGA, etc.). L'exemple suivant montre un enregistrement de rapport sur la cytosine :

```
chr2    24442367    +    18    0    CG    CGC
```

Pour générer un rapport sur le biais de structure M, mettez à l'option *--methylation-generate-mbias-report* la valeur « true » (activée). Ce rapport contient trois tableaux pour les données des bases à une extrémité, un tableau pour chaque contexte C et six tableaux pour les données des bases appariées. Chaque tableau est une suite d'enregistrements contenant un enregistrement par position de base de lecture. Par exemple, le premier enregistrement du tableau CHG contient les dénombrements de C méthylés (champ 2) et non méthylés (champ 3) qui se produisent dans la première position de base de lecture, mais seulement pour les lectures dont la première base est alignée à un emplacement CHG dans le génome. Chaque enregistrement d'un tableau comprend également le pourcentage de bases C méthylées (champ 4) et la somme des dénombrements de C méthylés et non méthylés (champ 5).

L'exemple suivant montre un enregistrement de biais de structure M pour la position de base de lecture 10 :

```
10    7335    2356    75.69    9691
```

Pour les ensembles de données contenant des lectures appariées qui se chevauchent, les rapports sur la cytosine et le biais de structure M ne signalent pas tout C de la deuxième lecture qui chevauche la première lecture. De plus, les coordonnées commençant par 1 sont utilisées pour les positions dans les deux rapports.

Pour faire correspondre les rapports bismark_methylation_extractor sur la cytosine et le biais de structure M généré par la version 0.19.0 de Bismark, mettez à l'option *--methylation-match-bismark* la valeur « true » (activée). L'ordre des enregistrements dans les rapports sur la cytosine de Bismark et de la plateforme DRAGEN peut être différent. Les rapports de la plateforme DRAGEN sont triés par position génomique.

Utilisation de Bismark pour la définition de la méthylation

L'approche recommandée pour la définition de la méthylation est d'automatiser les différents alignements nécessaires avec la plateforme DRAGEN, puis d'ajouter les étiquettes XM, XR et XG comme précédemment indiqué. Vous pouvez toutefois mettre à l'option *--enable-methylation-calling* la valeur « false » (désactivée) pour que la plateforme DRAGEN génère un fichier BAM distinct pour chacune des contraintes et des conversions du protocole de méthylation. Vous pouvez alors utiliser ces fichiers BAM avec un outil tiers de définition de la méthylation. Par exemple, vous pouvez modifier Bismark pour qu'il traite les lectures de ces fichiers BAM plutôt que d'exécuter Bowtie ou Bowtie2 à l'interne. Veuillez communiquer avec l'assistance technique d'Illumina si vous avec besoin d'aide avec ce processus.

Quand vous utilisez ce mode, une seule analyse par la plateforme DRAGEN génère plusieurs fichiers BAM dans le répertoire spécifié par l'option *--output-directory*. Chaque fichier contient les alignements dans le même ordre que les lectures d'entrée. Il n'est pas possible d'activer le tri ou le marquage des répétitions pour ces analyses. Les alignements comprennent les étiquettes MD et « /1 » ou « /2 », qui sont ajoutées aux noms des lectures appariées afin que ceux-ci soient compatibles avec Bismark. Ces fichiers BAM utilisent la convention de nommage suivante :

- Lectures à paire de bases unique – *output-directory/output-file-prefix.*{CT,GA}read {CT,GA}reference.bam
- Lectures appariées – *output-directory/output-file-prefix.*{CT,GA}read1{CT,GA}read2 {CT,GA}reference.bam,

où les valeurs de *output-directory* et de *output-file-prefix* sont spécifiées par les options correspondantes et où CT et GA correspondent aux conversions de bases indiquées dans le tableau ci-dessus.

Bismark n'a pas de mode directionnel avec compléments, mais vous pouvez traiter ce type d'échantillons grâce au mode non directionnel de Bismark en vous attendant à la production de très peu de bons alignements lors des analyses 1 et 2. C'est pourquoi, quand elle exécute un protocole non directionnel, la plateforme DRAGEN est automatiquement réglée pour investir moins d'efforts à la production d'alignements lors de ces analyses.

Chapitre 6 Préparation d'un génome de référence

Avant qu'un génome de référence puisse être utilisé par la plateforme DRAGEN, il doit être converti en format binaire personnalisé à partir du format FASTA pour pouvoir être utilisé par les composants matériels du système DRAGEN. Les options utilisées dans l'étape de prétraitement offrent un équilibre entre la performance et la qualité de la cartographie.

Le système DRAGEN est expédié avec les génomes de référence hg19 et GRCh37. Les deux génomes de référence sont préinstallés selon les paramètres recommandés pour les applications générales. Si vous trouvez que la performance et la qualité de la cartographie sont adéquates, vous pouvez probablement simplement travailler avec les génomes de référence fournis. Selon la longueur de vos lectures et d'autres aspects particuliers de votre application, vous pourriez être en mesure d'améliorer la qualité de la cartographie ou la performance en ajustant les options de prétraitement de la référence.

Aperçu de la table de hachage

La fonction de cartographie de la plateforme DRAGEN extrait plusieurs points d'ancrage qui se chevauchent (sous-séquences et k-mers) de chaque lecture et cherche ces points d'ancrage dans une table de hachage qui se trouve dans la mémoire de la carte PCIe afin d'identifier les emplacements dans le génome de référence où les points d'ancrage ont une correspondance. Les tables de hachage sont parfaites pour effectuer des recherches très rapides de correspondances exactes. La table de hachage de la plateforme DRAGEN doit être construite à partir d'un génome de référence donné à l'aide de l'option *dragen --build-hash-table* qui extrait plusieurs points d'ancrage qui se chevauchent à partir du génome de référence, qui les ajoute dans des enregistrements de la table de hachage et qui enregistre la table de hachage en fichier binaire.

Intervalle des points d'ancrage de la référence

La taille de la table de hachage de la plateforme DRAGEN est proportionnelle au nombre de points d'ancrage du génome de référence qui lui ont été ajoutés. Par défaut, un point d'ancrage qui commence à chacune des positions du génome de référence est ajouté à la table de hachage, c.-à-d. environ 3 milliards de points d'ancrage pour un génome humain. Cette valeur par défaut nécessite au moins 32 Go de mémoire sur la carte PCIe de la plateforme DRAGEN.

Pour utiliser des génomes non humains plus grands ou pour réduire la congestion de la table de hachage, il est possible de ne pas ajouter tous les points d'ancrage de référence en utilisant l'option *--ht-ref-seed-interval* pour spécifier un intervalle de référence moyen. La valeur d'intervalle par défaut pour ajouter tous les points d'ancrage est *--ht-ref-seed-interval 1*. Pour ajouter 50 % des points d'ancrage, vous pouvez utiliser l'option *--ht-ref-seed-interval 2*. L'intervalle d'ajout des points d'ancrage peut être un nombre non entier. Par exemple, l'option *--ht-ref-seed-interval 1.2* signifie que 83,3 % des bases sont ajoutées, avec une majorité d'intervalles d'une base et quelques intervalles de 2 bases pour obtenir un intervalle de 1,2 base en moyenne.

Taux d'occupation de table de hachage

Il s'agit d'une caractéristique des tables de hachage. Quand un espace leur a été attribué, mais que certains enregistrements sont vides, le taux d'occupation est inférieur à 100 %. Une certaine quantité d'espace vide est importante pour accéder rapidement à la table de hachage de la plateforme DRAGEN. Un taux d'occupation d'environ 90 % est une bonne limite supérieure. L'espace vide est important parce que les enregistrements sont placés pseudo-aléatoirement dans la table de hachage, ce qui entraîne un nombre d'enregistrements anormalement élevé à certains endroits. Ces régions de congestion peuvent être

particulièrement grandes quand le taux d'occupation s'approche de 100 %. Les demandes de la fonction de cartographie de la plateforme DRAGEN pour certains points d'ancrage peuvent alors être de plus en plus longues à exécuter.

Table de hachage et longueur de point d'ancrage

La table de hachage est remplie avec des points d'ancrage de référence ayant tous la même longueur. La longueur des points d'ancrage primaires est réglée avec l'option *--ht-seed-len* dont la valeur par défaut est 21.

La longueur maximale des points d'ancrage primaires est de 27 bases quand la table a une taille de 8 à 31,5 Go. Généralement, des points d'ancrage plus longs offrent une meilleure performance en termes de durée d'exécution, alors que des points d'ancrage plus courts sont meilleurs pour la qualité de la cartographie (taux de réussite et exactitude). Un point d'ancrage plus long a plus de chances d'être unique dans le génome de référence, ce qui permet d'accélérer la cartographie en éliminant le besoin de vérifier nombre de différents emplacements. Par contre, un point d'ancrage plus long a également plus de chances de chevaucher un écart de la référence (un variant ou une erreur de séquençage), ce qui fait échouer la cartographie d'une correspondance exacte (même si un autre point d'ancrage de la lecture pourrait quand même être cartographié). Il y a également moins de positions pour les points d'ancrage plus longs dans chaque lecture.

Les points d'ancrage plus longs sont plus utiles avec des lectures plus longues, car il y a plus de positions de point d'ancrage possibles, ce qui permet d'éviter les écarts.

Tableau 8 Recommandations relatives à la longueur de point d'ancrage

Valeur de *-ht-seed-len*	Longueur de lecture
21	100 à 150 pb
17 à 19	lectures plus courtes (36 pb)
27	250 pb et plus

Table de hachage et extensions de point d'ancrage

Comme il y a des séquences répétitives, certains points d'ancrage d'une longueur donnée peuvent correspondre à de nombreux emplacements dans le génome de référence. La plateforme DRAGEN utilise un mécanisme unique appelé l'extension de point d'ancrage pour réussir à cartographier les points d'ancrage ayant une fréquence aussi élevée. Quand le logiciel détermine qu'un point d'ancrage primaire se trouve à plusieurs emplacements de la référence, il étend le point d'ancrage d'un certain nombre de bases aux deux extrémités pour obtenir une séquence qui est plus unique dans la référence.

Par exemple, un point d'ancrage primaire de 21 bases peut être étendu de 7 bases à chaque extrémité pour obtenir un point d'ancrage étendu de 35 bases. Un point d'ancrage primaire de 21 bases peut correspondre à 100 emplacements dans la référence. Par contre, des extensions de 35 bases à ces 100 positions de point d'ancrage pourraient être divisées en 40 groupes de 1 à 3 points d'ancrage de 35 bases identiques. Les extensions itératives de point d'ancrage sont également prises en charge et sont automatiquement générées quand un grand ensemble de points d'ancrage primaires identiques contient différents sous-ensembles qui peuvent être plus facilement résolus à l'aide de différentes longueurs d'extension.

La longueur maximale des points d'ancrage étendus, équivalant par défaut à la longueur des points d'ancrage primaires plus 128, peut être modifiée avec l'option *--ht-max-ext-seed-len*. Par exemple, pour des lectures courtes, il est préférable de mettre une longueur maximale de points d'ancrage plus petite que la longueur de la lecture, car aucune correspondance ne peut être établie avec des extensions plus longues que la longueur de la lecture.

Il est également possible de régler à quel point les points d'ancrage seront étendus à l'aide des options suivantes (utilisation avancée) :

- *--ht-cost-coeff-seed-len*
- *--ht-cost-coeff-seed-freq*
- *--ht-cost-penalty*
- *--ht-cost-penalty-incr*

Il y a un équilibre entre la longueur des extensions et la fréquence des occurrences. La cartographie peut être effectuée plus rapidement en utilisant des extensions de point d'ancrage plus longues afin de réduire la fréquence d'occurrence des points d'ancrage. Ou la cartographie peut être plus exacte en évitant d'utiliser des extensions de point d'ancrage ou en limitant leur longueur, mais en devant tolérer les fréquences d'occurrences plus élevées que cela entraîne. Des extensions courtes peuvent améliorer la qualité de la cartographie en étant en mesure de placer les points d'ancrage entre les polymorphismes mononucléotidiques et en trouvant plus d'emplacements de cartographie possibles où peuvent être notés les alignements. Les paramètres d'extension et de fréquence de point d'ancrage par défaut favorisent fortement la qualité de la cartographie grâce à des extensions de point d'ancrage relativement courtes et des fréquences d'occurrences élevées.

Les valeurs par défaut des options de fréquence de point d'ancrage sont les suivantes :

Option	Valeur par défaut
--ht-cost-coeff-seed-len	1
--ht-cost-coeff-seed-freq	0,5
--ht-cost-penalty	0
--ht-cost-penalty-incr	0,7
--ht-max-seed-freq	16
--ht-target-seed-freq	4

Fréquence de point d'ancrage limite et cible

Un point d'ancrage primaire ou étendu peut correspondre à plusieurs emplacements dans le génome de référence. Toutes ces correspondances sont ajoutées dans la table de hachage, puis récupérées quand la fonction de cartographie de la plateforme DRAGEN cherche un point d'ancrage correspondant extrait d'une lecture. Les multiples positions de référence sont alors examinées et comparées pour générer une sortie de la fonction de cartographie avec alignement. Toutefois, *dragen* applique une limite sur le nombre de correspondances, ou la fréquence, de chaque point d'ancrage qui peut être réglée à l'aide de l'option *--ht-max-seed-freq*. Par défaut, la limite de fréquence est « 16 ». En pratique, quand le logiciel traite un point d'ancrage qui a une fréquence plus élevée, il l'étend jusqu'à en faire un point d'ancrage secondaire suffisamment long pour que la fréquence de tout point d'ancrage secondaire se trouve en deçà de la limite. Toutefois, si la limite de fréquence est dépassée par un point d'ancrage étendu au maximum, le point d'ancrage est rejeté et n'est pas ajouté à la table de hachage. Dans ce cas, *dragen* ajoute un seul enregistrement de fréquence élevée.

Cette limite de fréquence de point d'ancrage n'a pas tendance à affecter de manière importante la qualité de la cartographie de la plateforme DRAGEN et ce, pour deux raisons. Premièrement, puisque les points d'ancrage ne sont rejetés que quand l'extension échoue, seuls les points d'ancrage primaires ayant une fréquence extrêmement élevée, typiquement avec plusieurs milliers de correspondances, sont rejetés. Ces points d'ancrage ne sont pas très utiles pour la cartographie. Deuxièmement, il existe d'autres positions de point d'ancrage à vérifier dans une lecture donnée. Si une autre position de point d'ancrage est suffisamment unique pour retourner une ou plusieurs correspondances, la lecture peut quand même être

correctement cartographiée. Toutefois, si toutes les positions de point d'ancrage sont rejetées en raison d'une fréquence élevée, cela signifie souvent que l'ensemble de la lecture correspond assez bien à plusieurs positions de référence. Donc, même si la lecture était cartographiée, ce serait un choix arbitraire et elle aurait un score MAPQ très bas ou égal à zéro.

C'est pourquoi la limite de fréquence par défaut de 16 de l'option *--ht-max-seed-freq* fonctionne bien. Cette limite peut toutefois être diminuée ou augmentée et ce jusqu'à une valeur de 256. Une limite de fréquence plus élevée a tendance à augmenter un peu le nombre de lectures cartographiées (particulièrement des lectures courtes), mais ces lectures cartographiées additionnelles ont généralement un score MAPQ très bas ou égal à zéro. Ceci tend également à ralentir la cartographie par la plateforme DRAGEN, car un nombre proportionnellement élevé de cartographies possibles sont parfois analysées.

En plus d'une limite de fréquence, il est possible de spécifier une fréquence de point d'ancrage *cible* à l'aide de l'option *--ht-target-seed-freq*. Cette fréquence cible est utilisée quand des extensions sont générées pour des points d'ancrage primaires à fréquence élevée. Les longueurs des extensions sont choisies en favorisant les fréquences de point d'ancrage étendu qui sont proches de la fréquence cible. La valeur par défaut de l'option *--ht-target-seed-freq* est 4, ce qui entraîne un biais du logiciel envers la génération d'extensions de point d'ancrage plus courtes que la longueur nécessaire pour cartographier de manière unique les points d'ancrage.

Traitement des contigs de leurres

Le comportement de la plateforme DRAGEN en ce qui a trait au traitement des contigs de leurres dans la référence a changé depuis la version 2.6.

Depuis la plateforme DRAGEN 3.x, la fonction de construction de table de hachage détecte l'absence des contigs de leurres dans la référence et l'indique dans le fichier FASTA avant de construire la table de hachage. Le fichier de leurres se trouve sous /opt/edico/liftover/hs_decoys.fa. Si la référence ne contient aucun contig de leurres, alors la cartographie des lectures cartographiées aux contigs de leurres est artificiellement marquée comme supprimée dans le fichier BAM de sortie (puisque la référence originale ne contient pas de contig de leurres). Cela entraîne un taux de cartographie artificiellement plus bas. Toutefois, l'exactitude de la définition des variants est améliorée grâce au retrait des faux positifs causés par les lectures de leurres.

Illumina recommande l'utilisation de cette fonction par défaut. Vous pouvez néanmoins mettre à l'option *--htsuppress-decoys* la valeur « true » (activée) pour supprimer l'ajout de ces leurres à la table de hachage.

Le tableau ci-dessous décrit la différence de comportement entre les versions antérieures de la plateforme DRAGEN (jusqu'à 2.6) et les versions 3.x en ce qui a trait au traitement des contigs de leurres dans la fonction de construction de table de hachage :

Comportement de la plateforme DRAGEN	Version 2.6 et versions antérieures	Versions 3.x
La référence n'inclut pas de contigs de leurres (p. ex., GRCh37)	Les lectures de leurres sont incorrectement cartographiées ailleurs dans le génome à cause du manque de contigs dans la référence. • Le taux de cartographie est artificiellement élevé. • Définition de faux positifs dans des régions de bruit auxquelles les contigs de leurres sont incorrectement cartographiés.	La plateforme DRAGEN détecte automatiquement l'absence de contigs de leurres dans la référence et l'indique dans le fichier FASTA. • Le taux de cartographie est artificiellement bas, car les lectures de leurres cartographiés aux contigs de leurres sont artificiellement marquées comme supprimées dans le fichier BAM de sortie (puisque la référence originale ne contient pas de contig de leurres). • La définition de faux positifs est évitée grâce à l'ajout des contigs de leurres au processus. Par conséquent, cela améliore la définition de variants.
La référence inclut des contigs de leurres (p. ex., hs37d5)	Les lectures de leurres sont cartographiées aux contigs de leurres. • Le taux de cartographie est élevé. • Aucune définition de faux positif causée par la présence de lectures de leurres, car les lectures de leurres sont cartographiées au bon endroit.	Les lectures de leurres sont cartographiées aux contigs de leurres. • Le taux de cartographie est élevé. • Aucune définition de faux positif causée par la présence de lectures de leurres, car les lectures de leurres sont cartographiées au bon endroit.

Tables de hachage sensibles à la valeur du champ ALT

Pour activer la cartographie sensible à la valeur du champ ALT sur la plateforme DRAGEN, construisez le génome GRch38 (et les autres références avec contigs ALT) à partir d'un un fichier LiftOver en utilisant l'option *--ht-alt-liftover*. La fonction de construction de table de hachage classe chaque séquence de référence comme primaire ou secondaire selon le fichier LiftOver et place les séquences primaires avant les secondaires dans le fichier reference.bin. Les fichiers LiftOver de SAM des génomes hg38DH et hg19 sont fournis dans le dossier /opt/edico/liftover. L'option *--ht-alt-liftover* spécifie le chemin d'accès vers le fichier LiftOver pour la construction d'une table de hachage sensible à la valeur du champ ALT.

Pour retirer l'obligation d'utiliser un fichier LiftOver, mettez à l'option *--ht-alt-aware-validate* la valeur « false » (désactivée) lors de la construction des tables de hachage et quand vous exécuter la commande dragen.

Fichiers LiftOver personnalisés

Des fichiers LiftOver personnalisés peuvent être utilisés plutôt que ceux fournis avec la plateforme DRAGEN. Les fichiers LiftOver doivent être dans le format SAM, mais aucun en-tête n'est requis. Les champs SEQ et QUAL peuvent être omis (valeur « * »). Chaque enregistrement d'alignement devrait avoir un nom de séquence d'haplotype alternatif de référence dans le champ QNAME qui indique la valeur des champs RNAME et POS de son alignement LiftOver dans une séquence de référence de destination (habituellement, l'assemblage primaire).

Les alignements de compléments inversés sont indiqués par le bit « 0x10 » dans le champ FLAG. Les enregistrements marqués comme non cartographiés (« 0x4 ») ou secondaires (« 0x100 ») sont ignorés. La chaîne CIGAR peut inclure les bases coupées conservées ou non dans l'alignement, ce qui laisse des parties du contig ALT non alignées.

Une seule séquence de référence ne peut servir à la fois de contig ALT (qui apparaît dans le champ QNAME) et de destination LiftOver (qui apparaît dans le champ RNAME). Plusieurs contigs ALT peuvent s'aligner au même emplacement d'assemblage primaire. Plusieurs alignements peuvent également être fournis pour un seul contig ALT (les autres peuvent être marqués comme additionnels avec la valeur « 0x800 ») comme pour aligner une partie dans l'ordre normal et une autre dans l'ordre du complément inversé. Toutefois, chaque base d'un contig ALT ne reçoit qu'une image LiftOver, selon le premier enregistrement d'alignement dont la base est couverte par l'opération CIGAR M.

Les enregistrements SAM sans valeur dans le champ QNAME qui proviennent du génome de référence sont ignorés. Le même fichier LiftOver peut donc servir pour divers sous-ensembles de référence, mais une erreur survient si un alignement a une valeur dans le champ QNAME et aucune dans le champ RNAME.

Options de ligne de commande

Utilisez l'option *--build-hash-table* pour transformer un fichier FASTA de référence en table de hachage pour effectuer la cartographie par la plateforme DRAGEN. L'option prend un fichier FASTA (plusieurs séquences de référence qui sont concaténées) et un répertoire de sortie déjà existant en entrée pour générer l'ensemble de fichiers suivant :

reference.bin	Les séquences de référence, codées sur 4 bits par base. Comme des codes de 4 bits sont utilisés, la taille en octets est d'environ la moitié de celle du génome de référence. Entre les séquences de référence, N bases sont retranchées et des élargissements sont automatiquement insérés. Par exemple, hg19 a 3 137 161 264 bases dans 93 séquences. Celles-ci sont codées dans 1 526 285 312 octets, c.-à-d. 1,46 Go, où 1 Go équivaut à 1 Gio ou 2^{30} octets.
hash_table.cmp	La table de hachage compressée. La table de hachage est décompressée et utilisée par la fonction de cartographie de la plateforme DRAGEN pour chercher des points d'ancrage primaires de la longueur spécifiée par l'option *--ht-seed-len* et des points d'ancrage étendus de différentes longueurs.
hash_table.cfg	Une liste de paramètres et d'attributs pour la table de hachage, en format texte. Ce fichier fournit des renseignements importants sur le génome de référence et la table de hachage.
hash_table.cfg.bin	Une version binaire du fichier hash_table.cfg qui sert à la configuration des composants matériels de la plateforme DRAGEN.
hash_table_stats.txt	Un fichier texte contenant une liste complète des statistiques internes sur le hachage construit, notamment les pourcentages d'occupation de la table de hachage. Cette table n'est fournie qu'à titre informatif. Elle n'est pas utilisée par d'autres outils.

La commande de construction est utilisée comme suit :

```
dragen --build-hash-table true [options] --ht-reference <reference.fasta>
   --output-directory <outdir>
```

Les sections qui suivent fournissent des renseignements sur les options pour la construction d'une table de hachage.

Options d'entrée et de sortie

Les options *--ht-reference* et *--output-directory* sont obligatoires pour construire une table de hachage. L'option *--ht-reference* spécifie le chemin d'accès vers le fichier FASTA de référence, alors que l'option *--output-directory* spécifie le répertoire existant où les fichiers de sortie de table de hachage sont écrits. Illumina recommande d'organiser les constructions de table de hachage dans différents dossiers.

Comme pratique exemplaire, les noms de dossier devraient inclure tous les réglages des paramètres différents de ceux par défaut qui ont été utilisés pour générer la table de hachage que le dossier contient. Les noms de séquence du fichier FASTA de référence doivent être uniques.

Longueur de point d'ancrage primaire

L'option *--ht-seed-len* spécifie la longueur initiale en nucléotides des points d'ancrage du génome de référence qui sont ajoutés à la table de hachage. Lors de l'exécution, la fonction de cartographie extrait des points d'ancrage de cette longueur à partir de chaque lecture et recherche des correspondances exactes (sauf si la modification des points d'ancrage est activée) dans la table de hachage.

La longueur maximale de point d'ancrage primaire dépend de la taille de la table de hachage. La limite est k = 27 pour les tables de 16 à 64 Go, les tailles typiques pour le génome humain entier, ou k = 26 pour les tables de 4 à 16 Go.

La longueur minimale de point d'ancrage dépend principalement de la taille et de la complexité du génome de référence. Elle doit être suffisamment longue pour pouvoir déterminer avec unicité la plupart des positions de référence. Pour les références de génome humain entier, la construction de la table de hachage échoue généralement quand la valeur de k est inférieure à 16. La limite inférieure peut être plus basse pour des génomes plus courts ou être plus élevée pour des génomes moins complexes (plus répétitifs). Le seuil d'unicité de l'option *--ht-seed-len 16* pour un génome humain de 3,1 Gpb peut être compris intuitivement, car comme log4(3,1 G) ≈ 16, il est nécessaire d'avoir au moins 16 choix pour pouvoir distinguer 3,1 milliards de positions de référence à partir de 4 nucléotides.

Considérations relatives à l'exactitude

Pour que la cartographie de lectures réussisse, au moins un point d'ancrage primaire doit être une correspondance exacte (ou avec un seul polymorphisme mononucléotidique quand des points d'ancrage modifiés sont utilisés). Des points d'ancrage plus courts sont plus à même d'être cartographiés à la référence, car il est moins probable qu'ils chevauchent des variants ou des erreurs de séquençage et parce que plus de points d'ancrage peuvent être placés dans chaque lecture. Donc, en termes d'exactitude de la cartographie, les points d'ancrage plus petits sont généralement meilleurs.

Toutefois, les points d'ancrage très courts peuvent parfois réduire l'exactitude de la cartographie. Les points d'ancrage très courts vont souvent être cartographiés à plusieurs positions de référence, ce qui entraîne la prise en considération de plus de faux emplacements de cartographie par la fonction de cartographie. En raison de modèles imparfaits de mutations et d'erreurs par la notation d'alignement Smith-Waterman et par d'autres méthodes heuristiques, ces correspondances de bruit peuvent parfois être signalées. Les filtres de qualité de durée d'analyse comme *--Aligner.aln_min_score* peuvent régler les problèmes d'exactitude liés aux points d'ancrage très courts.

Considérations relatives à la durée d'exécution

Des points d'ancrage plus courts peuvent ralentir la cartographie, car ils peuvent être cartographiés à plus d'emplacements de référence, ce qui entraîne un traitement supplémentaire, comme des alignements Smith-Waterman, pour déterminer quel est le meilleur résultat. Cet effet se fait le plus sentir quand la longueur de point d'ancrage primaire est près du seuil d'unicité du génome de référence, c.-à-d. k = 16 pour le génome humain entier.

Considérations relatives à l'application

- Longueur des lectures – Généralement, les points d'ancrage plus courts sont appropriés aux lectures plus courtes et les points d'ancrage plus longs le sont pour les lectures plus longues. Dans une lecture courte, quelques positions de mésappariement (des variants ou des erreurs de séquençage) peuvent

découper la lecture en seulement des segments courts qui correspondent à la référence, de sorte que seul un point d'ancrage court peut être placé entre les différences et correspondre exactement à la référence. Par exemple, dans une lecture de 36 paires de bases, il ne faut qu'un polymorphisme mononucléotidique dans le milieu pour bloquer la correspondance de points d'ancrage de plus de 18 paires de bases avec la référence. Par contre, dans une lecture de 250 paires de bases, il faut 15 polymorphismes mononucléotidiques pour qu'il y ait une probabilité de plus de 0,01 % de bloquer la correspondance de points d'ancrage de 27 paires de bases.

- Lectures appariées – L'utilisation de lectures appariées peut permettre d'obtenir une bonne exactitude de cartographie avec des points d'ancrage plus longs. La plateforme DRAGEN se sert des renseignements des lectures appariées pour améliorer l'exactitude de la cartographie, notamment grâce à des analyses de récupération qui cherchent la fenêtre de référence attendue quand il n'y a des points d'ancrage cartographiés que pour une des lectures appariées dans une région de référence donnée. Donc, pour les lectures appariées, il y a essentiellement deux fois plus d'occasions de rechercher une correspondance exacte de point d'ancrage pour trouver de bons alignements.
- Taux d'erreurs et de variants – Quand les différences entre la lecture et la référence sont plus fréquentes, il peut être nécessaire d'utiliser des points d'ancrage plus courts pour être en mesure de les placer entre les positions des différences d'une lecture donnée afin de trouver une correspondance exacte à la référence.
- Exigence de pourcentage de cartographie – Si l'application nécessite qu'un pourcentage élevé des lectures soient cartographiées (même avec un score MAPQ bas), des points d'ancrage courts peuvent être utiles. Il est plus probable que certaines lectures qui ne correspondent bien à la référence à aucun emplacement puissent être cartographiées à l'aide de points d'ancrage courts en trouvant des correspondances partielles à la référence.

Longueur maximale de point d'ancrage

L'option *--ht-max-ext-seed-len* limite la longueur des points d'ancrage étendus qui sont ajoutés dans la table de hachage. Les points d'ancrage primaires (longueur spécifiée par l'option *--ht-seed-len*) qui correspondent à plusieurs positions de référence peuvent être étendus pour obtenir une mise en correspondance plus unique, ce qui peut être nécessaire pour cartographier les points d'ancrage en deçà de la fréquence maximale d'occurrences (option *--ht-max-seed-freq*).

Avec une longueur de point d'ancrage primaire k, la longueur maximale du point d'ancrage peut être entre k et k+128. La valeur par défaut est la limite supérieure de k+128.

Quand limiter l'extension de point d'ancrage

Il est recommandé d'utiliser l'option *--ht-max-ext-seed-len* pour les lectures courtes, c.-à-d. celles de moins de 50 paires de bases. Dans ces situations, il est utile de limiter l'extension de point d'ancrage à la longueur de la lecture moins une petite marge, c.-à-d. de 1 à 4 paires de bases. Par exemple, avec des lectures de 36 paires de bases, il est approprié de mettre à l'option *--ht-max-ext-seed-len* une valeur de « 35 ». Ceci permet d'assurer que la fonction de construction de table de hachage ne prévoit pas utiliser d'extensions de points d'ancrage plus longs que la lecture, ce qui entraînerait l'échec des extensions et de la cartographie lors de l'exécution pour les points d'ancrage qui pourraient être placés dans la lecture s'ils avaient des extensions plus courtes.

Bien que la limite d'extension de point d'ancrage puisse être également appliquée aux lectures plus longues, p. ex. en mettant à l'option *--ht-max-ext-seed-len* la valeur « 99 » pour des lectures de 100 paires de bases, ce ne serait que peu utile, car les points d'ancrage ne sont étendus que modérément dans tous les cas. Même avec la limite par défaut de k+128, les points d'ancrage individuels ne sont étendus que pour

atteindre une longueur en deçà de la fréquence maximale d'occurrences (option *--ht-max-seed-freq*), tout au plus de quelques bases supplémentaires pour s'approcher de la fréquence d'occurrences cible (option *--ht-target-seed-freq*), ou pour éviter qu'il y ait trop d'étapes incrémentales d'extension.

Fréquence maximale d'occurrences

L'option *--ht-max-seed-freq* établit une limite ferme du nombre d'occurrences de points d'ancrage (emplacements du génome de référence) qui peuvent être ajoutés pour tout point d'ancrage primaire ou étendu. Si un point d'ancrage primaire est cartographié à un nombre de positions de référence supérieur à cette limite, il doit être étendu suffisamment pour que les points d'ancrage étendus se subdivisent en plus petits groupes de points d'ancrage identiques en deçà de la limite. Si un groupe de points d'ancrage de référence est plus grand que cette limite, même à la longueur maximale de point d'ancrage étendu (*--ht-max-ext-seed-len*), les positions de référence ne sont pas ajoutées dans la table de hachage. Dans ce cas, *dragen* ajoute un seul enregistrement de fréquence élevée.

La fréquence maximale d'occurrences peut avoir une valeur de 1 à 256. Toutefois, si cette valeur est trop basse, la construction de la table de hachage peut échouer, car elle nécessitera trop d'extensions de point d'ancrage. La valeur minimale utile pour une référence de génome humain complet est de 8 si les valeurs par défaut des autres options sont utilisées.

Considérations relatives à l'exactitude

Généralement, une fréquence maximale d'occurrences plus élevée entraîne une meilleure réussite de la cartographie. Il y a deux raisons pour ceci. Premièrement, quand la limite est plus élevée, moins de positions de référence ne pouvant être cartographiées en deçà de celle-ci sont rejetées. Deuxièmement, une limite plus élevée permet des extensions de point d'ancrage plus courtes, ce qui améliore les probabilités de correspondance exacte d'un point d'ancrage sans chevauchement de variants ou d'erreurs de séquençage.

Toutefois, comme pour les points d'ancrage très courts, permettre des dénombrements d'occurrences élevés peut parfois nuire à l'exactitude de la cartographie. La plupart des occurrences de point d'ancrage d'un grand groupe ne sont pas de véritables emplacements de cartographie. Parfois, une de ces occurrences de bruit est signalée en raison de modèles de notation imparfaits. De plus, la fonction de cartographie limite le nombre total d'emplacements de référence qu'elle prend en considération. Permettre des dénombrements d'occurrences élevés peut potentiellement empêcher l'évaluation de la meilleure correspondance.

Considérations relatives à la durée d'exécution

Des fréquences maximales d'occurrences élevées peuvent allonger la cartographie des lectures, car la cartographie des points d'ancrage trouve plus d'emplacements de référence, ce qui entraîne plus de traitement, comme des alignements Smith-Waterman, pour déterminer quel est le meilleur résultat.

Options de fichier LiftOver sensible à la valeur du champ ALT

Consultez la section *Tables de hachage sensibles à la valeur du champ ALT* à la page 145 pour en savoir plus sur la construction d'un fichier LiftOver personnalisé.

- *--ht-alt-liftover*

  L'option *--ht-alt-liftover* spécifie le chemin d'accès vers le fichier LiftOver pour la construction d'une table de hachage sensible à la valeur du champ ALT. Cette option est obligatoire pour la construction à partir d'une référence avec des contigs ALT. Les fichiers LiftOver de SAM pour hg38DH et hg19 sont fournis dans le répertoire /opt/edico/liftover.

- *--ht-alt-aware-validate*

L'utilisation d'un fichier LiftOver est obligatoire lors de la construction de tables de hachage à partir d'une référence qui contient des contigs ALT. Pour désactiver cette exigence, mettez à l'option *--ht-alt-aware-validate* la valeur « false » (désactivée).

- *--ht-decoys*

  Le logiciel de la plateforme DRAGEN détecte automatiquement l'utilisation des références hg19 et hg38 et ajoute des leurres à la table de hachage quand il n'en trouve pas dans le fichier FASTA. Utilisez l'option *--ht-decoys* pour spécifier le chemin d'accès vers un fichier de leurres. Le chemin d'accès par défaut est /opt/edico/liftover/hs_decoys.fa.

- *--ht-suppress-decoys*

  Utilisez l'option *--ht-suppress-decoys* pour supprimer l'utilisation du fichier de leurres lors de la construction de la table de hachage.

Options du logiciel de la plateforme DRAGEN

- *--ht-num-threads*

  L'option *--ht-num-threads* détermine le nombre maximal de fils de travail de l'unité centrale qui sont utilisés pour accélérer la construction de la table de hachage. La valeur par défaut de cette option est 8, avec un maximum de 32 fils permis.

  Si votre serveur prend en charge l'exécution de plus de fils, il est recommandé d'utiliser le maximum de fils possible. Par exemple, les serveurs de la plateforme DRAGEN contiennent 24 cœurs avec technologie Hyperthread, donc une valeur de 32 devrait être utilisée. Quand vous utilisez une valeur plus élevée, vous devez également ajuster la valeur de l'option *--ht-max-table-chunks*. Les serveurs sont dotés de 128 Go de mémoire disponible.

- *--ht-max-table-chunks*

  L'option *--ht-max-table-chunks* contrôle la quantité de mémoire utilisée pendant la construction de la table de hachage en limitant le nombre de blocs de table de hachage d'environ 1 Go se trouvant simultanément en mémoire. Chaque bloc additionnel utilise environ deux fois sa taille (environ 2 Go) dans la mémoire système pendant la construction.

  La table de hachage est divisée en un nombre entier à la puissance deux de blocs indépendants de taille fixe, X, qui dépend de la taille de la table de hachage, dans l'intervalle 0,5 Go < X ≤ 1 Go. Par exemple, une table de hachage de 24 Go contient 32 blocs indépendants de 0,75 Go qui peuvent être construits par des fils parallèles s'il y a suffisamment de mémoire, alors qu'une table de hachage de 16 Go contient 16 blocs indépendants de 1 Go.

  La valeur par défaut de l'option *--ht-max-table-chunks* est égale à la valeur de l'option *--ht-num-threads*, avec une valeur minimale par défaut de 8. Il est logique que les valeurs de ces deux options soient les mêmes, car la construction d'un bloc de table de hachage nécessite un fil pour l'exécution et l'équivalent d'un bloc en mémoire. Toutefois, il y a certains avantages relatifs à la vitesse de construction à mettre à l'option *--ht-max-table-chunks* une valeur plus élevée que celle de l'option *--ht-num-threads*, et vice-versa.

Options relatives à la taille

- *--ht-mem-limit* – Limite de mémoire

  L'option *--ht-mem-limit* permet de déterminer la taille de la table de hachage générée en spécifiant la mémoire disponible de la carte DRAGEN pour la table de hachage et le génome de référence codé. L'option *--ht-mem-limit* a une valeur par défaut de 32 Go quand la taille du génome de référence est près de celle d'un génome humain entier, ou une taille généreuse pour les références plus petites.

Habituellement, il n'est pas nécessaire de modifier ces valeurs par défaut.

- *--ht-size* – Taille de la table de hachage

  Cette option spécifie la taille de la table de hachage à générer plutôt que de calculer une taille de table appropriée en fonction de la taille du génome de référence et de la mémoire disponible (option *--ht-mem-limit*). Il est recommandé d'utiliser les valeurs par défaut pour déterminer la taille de la table. Sinon, il est recommandé d'utiliser l'option *--ht-mem-limit*.

Options relatives à l'ajout de points d'ancrage

- *--ht-ref-seed-interval* – Intervalle entre les points d'ancrage

  L'option *--ht-ref-seed-interval* définit l'intervalle entre les positions des points d'ancrage dans le génome de référence qui sont ajoutés à la table de hachage. Un intervalle de 1 (valeur par défaut) signifie que chaque position de point d'ancrage est ajoutée, un intervalle de 2 signifie que 50 % des positions sont ajoutées, etc. Les nombres non entiers sont pris en charge, c.-à-d. qu'avec un intervalle de 2,5, 40 % des positions sont ajoutées.

  Les points d'ancrage d'un génome humain entier de référence peuvent facilement être tous ajoutés à des cartes DRAGEN dotées de 32 Go de mémoire. Si un génome de référence beaucoup plus grand est utilisé, modifiez la valeur de cette option.

- *--ht-soft-seed-freq-cap* et *--ht-max-dec-factor* – Limite souple de fréquence et facteur maximal d'élimination pour la réduction du nombre de points d'ancrage.

  La réduction du nombre de points d'ancrage est une technique expérimentale visant à améliorer la performance de la cartographie dans des régions à fréquence élevée. Quand les points d'ancrage primaires ont une fréquence plus élevée que la limite indiquée par l'option *--ht-soft-seed-freq-cap*, seule une partie des positions de point d'ancrage sera ajoutée afin de rester en deçà de cette limite. L'option *--ht-max-dec-factor* spécifie un facteur maximal par lequel le nombre de points d'ancrage peut être réduit. Par exemple, l'option *--ht-max-dec-factor 3* permet de conserver au moins un tiers des points d'ancrage initiaux. L'option *--ht-max-dec-factor 1* désactive la réduction du nombre de points d'ancrage.

  Les points d'ancrage sont éliminés selon des règles précises afin d'éviter des écarts trop grands entre les points d'ancrage. Le principe est que la réduction du nombre de points d'ancrage permet d'obtenir une couverture de points d'ancrage cartographiés dans des régions à fréquence élevée où la fréquence maximale d'occurrences aurait normalement été dépassée. La réduction du nombre de points d'ancrage permet également d'avoir des extensions de point d'ancrage plus courtes, ce qui aide aussi à réussir la cartographie. Selon les tests effectués à ce jour, il n'a pas été démontré que la réduction du nombre de points d'ancrage était supérieure aux autres méthodes d'optimisation de l'exactitude.

- *--ht-rand-hit-hifreq* et *--ht-rand-hit-extend* – Occurrence aléatoire de l'échantillon avec enregistrement HIFREQ et EXTEND

  Quand un enregistrement HIFREQ ou EXTEND est ajouté à la table de hachage, il représente un grand ensemble d'occurrences de référence pour un certain point d'ancrage. La fonction de construction de table de hachage peut également choisir aléatoirement un représentant de cet ensemble et l'ajouter à cet enregistrement HIT avec l'enregistrement HIFREQ ou EXTEND.

  Les occurrences aléatoires d'échantillon fournissent des alignements différents qui sont très utiles pour l'estimation du score MAPQ avec exactitude pour les alignements qui sont signalés. Ils ne sont pas utilisés autrement que dans ce contexte pour indiquer les positions d'alignement, car ceci entraînerait un biais dans la couverture des emplacements qui ont été sélectionnés pendant la construction de la table de hachage.

Pour inclure une occurrence de l'échantillon, mettez à l'option *--ht-rand-hit-hifreq* la valeur 1. L'option *--ht-rand-hit-extend* correspond au dénombrement minimal d'occurrences avant l'extension nécessaire pour inclure une occurrence de l'échantillon. Mettre la valeur à zéro désactive cette possibilité. Il n'est *pas* recommandé de modifier ces options.

Contrôle de l'extension des points d'ancrage

L'extension des points d'ancrage de la plateforme DRAGEN est dynamique et s'applique au besoin à certains k-mers qui correspondent à trop d'emplacements de référence. Les points d'ancrage sont étendus par incréments de 2 à 14 bases (toujours un nombre pair), d'une longueur de point d'ancrage primaire jusqu'à une longueur d'extension maximale. Les bases sont ajoutées symétriquement lors de chaque étape de l'extension et déterminent la prochaine incrémentation d'extension, le cas échéant.

Il y a un arbre d'extension de point d'ancrage potentiellement complexe associé à chaque fréquence élevée de points d'ancrage. Chacun des arbres complets est généré pendant la construction de la table de hachage et un chemin est tracé à partir de la racine par les étapes itératives d'extension lors de la cartographie du point d'ancrage. La fonction de construction de table de hachage utilise un algorithme de programmation dynamique pour chercher l'arbre d'extension de point d'ancrage optimal parmi tous les arbres possibles en se servant d'une fonction de coût qui établit un équilibre entre l'exactitude et la vitesse de la cartographie. La fonction de coût est définie à l'aide des options suivantes :

- *--ht-target-seed-freq* – Fréquence d'occurrences cible

  L'option *--ht-target-seed-freq* définit le nombre idéal d'occurrences par point d'ancrage qui est ciblé par l'extension de point d'ancrage. Des valeurs plus élevées produisent des extensions de point d'ancrage finales plus courtes et moins nombreuses, car des points d'ancrage plus courts tendent à correspondre à plus de positions de référence.

- *--ht-cost-coeff-seed-len* – Coefficient de coût pour la longueur de point d'ancrage

  L'option *--ht-cost-coeff-seed-len* attribue un élément de coût à chaque base ajoutée lors d'une extension de point d'ancrage. Ces bases additionnelles sont considérées comme un coût, car des points d'ancrage plus longs ont plus de risque de chevaucher des variants ou des erreurs de séquençage et de ne pouvoir être cartographiés. Des valeurs plus élevées entraînent la production d'extensions finales de point d'ancrage plus courtes.

- *--ht-cost-coeff-seed-freq* – Coefficient de coût pour la fréquence d'occurrences

  L'option *--ht-cost-coeff-seed-freq* attribue un élément de coût à la différence entre la fréquence d'occurrences cible et le nombre d'occurrences ajoutées pour un point d'ancrage. Des valeurs plus élevées entraînent principalement une plus grande extension des points d'ancrage à fréquence élevée pour que leur fréquence se rapproche de la fréquence cible.

- *--ht-cost-penalty* – Coût supplémentaire pour l'extension de point d'ancrage

  L'option *--ht-cost-penalty* attribue un coût fixe pour l'extension de point d'ancrage au-delà de la longueur de point d'ancrage primaire. Une valeur plus élevée entraîne une réduction du nombre d'extensions de point d'ancrage. Les tests actuels ont montré qu'une valeur de zéro (0) est appropriée pour ce paramètre.

- *--ht-cost-penalty-incr* – Incrément de coût par étape d'extension

  L'option *--ht-cost-penalty-incr* attribue un coût récurrent à chaque étape incrémentale d'extension de point d'ancrage, de la longueur de point d'ancrage primaire à la longueur finale du point d'ancrage après l'extension. Un grand nombre d'étapes entraîne un coût plus élevé, car l'extension en plusieurs petites étapes nécessite plus d'espace dans la table de hachage pour les enregistrements EXTEND intermédiaires, en plus d'allonger considérablement la durée d'exécution des extensions. Une valeur plus élevée produit des arbres d'extension de point d'ancrage avec moins de nœuds.

Les étapes sont donc plus grandes et moins nombreuses pour passer du nœud racine, la longueur de point d'ancrage primaire, aux nœuds terminaux, les longueurs de points d'ancrage étendus.

Tables de hachage propres à un pipeline

Quand vous construisez une table de hachage, la plateforme DRAGEN configure par défaut les options pour le traitement du séquençage de l'ADN. Pour analyser des données de séquençage de l'ARN, vous devez construire une table de hachage de séquençage de l'ARN à l'aide de l'option *--ht-build-rna-hashtable true*. Pour analyser des alignements de séquençage de l'ARN, consultez le répertoire *--output-directory* initial et non les sous-répertoires automatiquement générés.

Dans le cas du pipeline VNC, la table de hachage doit être construite avec l'option *--enable-cnv* avec la valeur « true » (activée), ce qui génère une table de hachage additionnelle de k-mers utilisée par l'algorithme relatif aux VNC. Illumina recommande de toujours utiliser l'option *--enable-cnv*, au cas où vous voudriez effectuer la définition de VNC en utilisant la même table de hachage que pour la cartographie et l'alignement.

Les analyses de méthylation de la plateforme DRAGEN nécessitent la construction d'une paire de tables de hachages spéciales dans lesquelles les bases de référence sont converties de C en T dans l'une et de G en T dans l'autre. Quand vous exécutez la génération de table de hachage avec l'option *--ht-methylated*, ces conversions sont effectuées automatiquement. Ces tables de hachage avec conversions sont générées dans une paire de sous-répertoires du répertoire cible qui est spécifié à l'aide de l'option *--output-directory*. Les sous-répertoires sont nommés CT_converted et GA_converted, ce qui correspond aux conversions automatiques des bases. Quand vous utilisez ces tables de hachage pour des analyses d'alignements méthylés, consultez le répertoire *--output-directory* initial et non les sous-répertoires générés automatiquement.

Comme ces conversions de bases suppriment énormément de renseignements des tables de hachage, il pourrait être nécessaire de modifier les paramètres de la table de hachage pour qu'ils soient différents de ceux d'une construction de table de hachage conventionnelle. Les options suivantes sont recommandées pour la construction de tables de hachage pour les mammifères :

```
dragen --build-hash-table=true --output-directory $REFDIR \
   --ht-reference $FASTA --ht-max-seed-freq 16 \
   --ht-seed-len 27 --ht-num-threads 40 --ht-methylated=true
```

Chapitre 7 Outils et utilitaires

Conversion des données en format BCL d'Illumina

Le format BCL est le format natif de sortie des systèmes de séquençage d'Illumina. Il est composé d'un répertoire contenant plusieurs fichiers de données et de métadonnées. Comme les fichiers de données sont organisés selon le schéma de Flow Cell du séquenceur, la conversion de ces données en fichiers FASTQ séparés par échantillon est une opération intensive qui peut prendre beaucoup de temps.

La plateforme DRAGEN offre une mise en œuvre du logiciel de conversion qui est rapide et qui prend en charge l'accélération matérielle sur la plateforme DRAGEN. Pour effectuer cette conversion, utilisez les options *--bcl-input-directory <BCL_ROOT>*, *--output-directory <DIR>* et *--bcl-conversion-only true*.

La mise en œuvre de la conversion en format BCL de la plateforme DRAGEN est conçue pour produire des fichiers FASTQ qui correspondent au format de sortie du logiciel bcl2fastq2 v2.20 d'Illumina. La plateforme DRAGEN prend en charge la plupart des fonctionnalités du logiciel bcl2fastq2, notamment le démultiplexage d'échantillons par code à barres avec tolérance de mésappariement facultative, le masquage ou le retranchement de séquences d'adaptateur avec une stringence de correspondance réglable et l'étiquetage et le retranchement de séquences d'IMU. Les paramètres de feuille d'échantillons qui ne sont pas pris en charge par la plateforme DRAGEN comprennent FindAdapterWithIndels, CreateFastqForIndexReads et ReverseComplement. La plateforme DRAGEN ne prend pas en charge l'option de ligne de commande no-lane-splitting.

Options de ligne de commande

Les options obligatoires pour la conversion des données en format BCL se trouvent dans l'exemple de commande ci-dessous :

```
dragen --bcl-conversion-only --bcl-input-directory <…> --output-directory
   <…>
```

Les options supplémentaires suivantes peuvent être spécifiées dans la ligne de commande :

- *--sample-sheet* – Spécifie le chemin d'accès vers le fichier SampleSheet.csv. L'option *--sample-sheet* est facultative si le fichier SampleSheet.csv se trouve dans le répertoire *--bcl-input-directory*.
- *--strict-mode* – Si la valeur de cette option est « true » (activée), la commande *dragen* s'arrêtera si des fichiers sont manquants. La valeur par défaut est « false » (désactivée).
- *--first-tile-only* – Si la valeur de cette option est « true » (activée), la commande *dragen* ne convertit que le premier fichier d'entrée (à des fins de test et de débogage). La valeur par défaut est « false » (désactivée).
- *--bcl-only-lane <#>* – La commande ne convertit que la ligne spécifiée dans la commande de conversion.
- *-f* – Conversion vers le fichier de sortie même si celui-ci existe déjà (forcer).
- *--bcl-use-hw false* – Pour ne pas utiliser l'accélération FPGA de la plateforme DRAGEN lors de la conversion en format BCL.

Le répertoire racine BCL d'entrée et le répertoire de sortie doivent être spécifiés. Le chemin d'accès d'entrée spécifié n'est pas un répertoire BaseCalls, il se trouve trois niveaux plus haut et devrait notamment contenir les fichiers et répertoires suivants :

```
Config\
Data\
Logs\
```

```
runParameters.xml
RunInfo.xml
```

Le répertoire où les fichiers FASTQ de sortie seront stockés est spécifié à l'aide de l'option *--output-dir*.

Options de feuille d'échantillons

En plus des options de ligne de commande, qui contrôlent le comportement de la conversion en format BCL, la feuille d'échantillons accepte les paramètres de la section [Settings] (Paramètres) du fichier de configuration spécifiant comment traiter les échantillons. Les paramètres de la feuille d'échantillons pour la conversion en format BCL sont les suivants :

Option	Valeur par défaut	Valeurs	Description
AdapterBehavior	trim	trim, mask	Indique si l'adaptateur doit être retranché ou masqué.
AdapterRead1	Aucune	Séquence d'adaptateur de la lecture 1 contenant A, C, G ou T	Séquence à retrancher ou à masquer à la fin de la lecture 1.
AdapterRead2	Aucune	Séquence d'adaptateur de la lecture 2 contenant A, C, G ou T	Séquence à retrancher ou à masquer à la fin de la lecture 2.
AdapterStringency	0,9	Valeur variant entre « 0,5 » et « 1,0 »	La rigueur pour faire correspondre la lecture avec l'adaptateur à l'aide de l'algorithme de fenêtre coulissante.
BarcodeMismatchesIndex1	1	0, 1 ou 2	Le nombre de mésappariements permis entre la première lecture d'index et la séquence d'indexage.
BarcodeMismatchesIndex2	1	0, 1 ou 2	Le nombre de mésappariements permis entre la seconde lecture d'index et la séquence d'indexage.
MinimumTrimmedReadLength	La plus petite valeur entre « 35 » et la longueur de la plus petite lecture non indexée.	Entre « 0 » et longueur de la plus petite lecture non indexée.	Les lectures précédemment retranchées deviennent masquées à partir de ce point.

Option	Valeur par défaut	Valeurs	Description
MaskShortReads	La petite valeur entre « 22 » et MinimumTrimmedReadLength.	Entre « 0 » et MinimumTrimmedReadLength	Les lectures précédemment retranchées sont complètement masquées.
OverrideCycles	Aucune	Y : Indique une lecture de séquençage I : Indique une lecture d'indexage U : Indique une longueur d'identifiants moléculaires uniques (IMU) à retrancher de la lecture	La chaîne utilisée pour spécifier les cycles relatifs aux identifiants moléculaires uniques (IMU) et les cycles masqués d'une lecture.

Les valeurs des éléments masqués de OverrideCycles sont séparés par des points-virgules. Par exemple :

```
OverrideCycles,N1Y150;I8;I7N1;Y141U10
```

La plateforme DRAGEN prend maintenant en charge le traitement flexible d'identifiants moléculaires uniques (IMU) durant la conversion au format BCL afin d'être compatible avec un plus grand nombre de tests tiers, y compris les séquences d'IMU dans les lectures d'index et les régions d'IMU multiples dans une lecture. Les séquences d'IMU sont retranchées des séquences de lectures de fichier FASTQ et placées dans l'identifiant de la séquence pour chaque lecture, en guise de référence.

Les exemples suivants montrent des paramètres OverrideCycles à partir de lectures 2x151 :

OverrideCycles,U7N1Y143;I8;I8;U7N1Y143	Des IMU non aléatoires sont utilisés par Illumina dans les produits suivants : • TruSight Oncology 170 RUO • TruSight Oncology 500 RUO • IDT pour Illumina – Ancres d'index avec IMU
OverrideCycles,U5N3Y143;I8;I8; U5N3Y143	Un IMU est composé des cinq premières paires de bases de chaque lecture génomique qui sont liées par trois paires de bases d'une séquence ignorée.
OverrideCycles,Y151;I8;U10;Y151	La lecture d'index 2 est un IMU.
OverrideCycles,Y151;I8U10;I8;Y151	La lecture d'index 1 contient à la fois un index et un IMU.
OverrideCycles,Y151;I8;I8;U10N12Y127	Un IMU se trouve au début de la lecture 2, joint à une séquence de liaison de longueur 12.

Sortie d'indicateurs du format BCL

Les indicateurs de conversion en format BCL de la plateforme DRAGEN sont produits dans le sous-dossier de sortie Reports/. Les renseignements fournis comprennent les indicateurs de démultiplexage, les indicateurs de commutation d'index (pour les index doubles uniques seulement) et les codes à barres inconnus les plus fréquents de chaque ligne.

Surveillance de l'intégrité du système

Quand vous allumez votre système DRAGEN, un programme démon (*dragen_mond*) démarre pour surveiller la carte et détecter les problèmes de matériel. Ce programme démon démarre également quand votre système DRAGEN est installé ou mis à jour. L'utilité principale de ce programme démon est de surveiller la température du processeur de la plateforme Bio-IT DRAGEN et d'arrêter la plateforme quand la température dépasse un seuil établi.

Pour lancer, arrêter ou redémarrer manuellement le programme de surveillance, exécutez la commande suivante en tant qu'utilisateur racine :

```
sudo service dragen_mond [stop|start|restart]
```

Par défaut, le programme de surveillance fait une vérification pour détecter les problèmes de matériel chaque minute et enregistre la température dans le journal chaque heure.

Le fichier /etc/sysconfig/dragen_mond spécifie les options de ligne de commande utilisées pour démarrer le programme *dragen_mond* quand la commande de service est exécutée. Modifiez le fichier DRAGEN_MOND_OPTS pour modifier les options par défaut. Par exemple, le délai de vérification est de 30 secondes et le délai de l'enregistrement est de 2 heures quand les options ci-dessous sont spécifiées :

```
DRAGEN_MOND_OPTS="-d -p 30 -l 7200"
```

L'option *-d* est requise pour exécuter le programme de surveillance en tant que programme démon.

Les options de ligne de commande de *dragen_mond* sont les suivantes :

Option	Description
-m --swmaxtemp <n>	La température maximale d'alarme logicielle (en degrés Celsius). La valeur par défaut est « 85 ».
-i --swmintemp <n>	La température minimale d'alarme logicielle (en degrés Celsius). La valeur par défaut est « 75 ».
-H --hwmaxtemp <n>	La température maximale d'alarme matérielle (en degrés Celsius). La valeur par défaut est « 100 ».
-p --polltime <n>	Le délai entre les vérifications d'état de la puce (en secondes). La valeur par défaut est « 60 ».
-l --logtime <n>	L'enregistrement de la température de FPGA chaque n secondes. La valeur par défaut est « 3 600 ». Cette valeur doit être un multiple du délai entre les vérifications.
-d --daemon	Détacher et exécuter en tant que programme démon.
-h --help	Affiche l'aide et s'arrête.
-V --version	Affiche la version et quitte.

Pour afficher la température actuelle du processeur de la plateforme DRAGEN Bio-IT, servez-vous de la commande *dragen_info -t*. Cette commande ne peut être lancée si le programme *dragen_mond* n'est pas en cours d'exécution.

```
% dragen_info -t
FPGA Temperature: 42C  (Max Temp: 49C, Min Temp: 39C)
```

Journaux

Tous les événements relatifs au matériel sont enregistrés dans le répertoire /var/log/messages et le fichier /var/log/dragen_mond.log. L'exemple suivant montre une alarme relative à la température dans /var/log/messages :

```
Jul 16 12:02:34 komodo dragen_mond[26956]: WARNING: FPGA software over temperature alarm has been
```

```
triggered -- temp threshold: 85 (Chip status: 0x80000001)
 Jul 16 12:02:34 komodo dragen_mond[26956]: Current FPGA temp: 86, Max temp: 88, Min temp: 48
 Jul 16 12:02:34 komodo dragen_mond[26956]: All dragen processes will be stopped until alarm clears
 Jul 16 12:02:34 komodo dragen_mond[26956]: Terminating dragen in process 1510 with SIGUSR2 signal
```

Par défaut, la température est enregistrée chaque heure dans le fichier /var/log/dragen_mond.log :

```
 Aug 01 09:16:50 Setting FPGA hardware max temperature threshold to 100
 Aug 01 09:16:50 Setting FPGA software max temperature threshold to 85
 Aug 01 09:16:50 Setting FPGA software min temperature threshold to 75
 Aug 01 09:16:50 FPGA temperatures will be logged every 3600 seconds
 Aug 01 09:16:50 Current FPGA temperature is 52 (Max temp = 52, Min temp = 52)
 Aug 01 10:16:50 Current FPGA temperature is 53 (Max temp = 56, Min temp = 49)
 Aug 01 11:16:50 Current FPGA temperature is 54 (Max temp = 56, Min temp = 49)
```

Si la plateforme DRAGEN est en cours d'exécution quand une alarme de température se déclenche, le message suivant s'affiche dans la fenêtre du terminal du processus de la plateforme DRAGEN :

```
************************************************************
**  Received external signal -- aborting dragen.          **
**  An issue has been detected with the dragen card.      **
**  Check /var/log/messages for details.                  **
**                                                        **
**  It may take up to a minute to complete shutdown.      **
************************************************************
```

Si ce message s'affiche, arrêtez l'exécution du logiciel de la plateforme DRAGEN. Prenez les mesures suivantes pour atténuer l'état de surchauffe de la carte :

- Assurez-vous qu'il y a une circulation d'air suffisante autour de la carte. Pensez à déplacer la carte dans un logement où il y a plus de circulation d'air, à ajouter un autre ventilateur ou à augmenter la vitesse du ventilateur.
- Laissez à la carte plus d'espace dans la boîte. S'il y a des logements PCIe libres, déplacez la carte de manière à ce qu'elle se trouve entre deux logements libres.

Communiquez avec l'assistance technique d'Illumina si n'arrivez pas à résoudre le problème lié à l'alarme de température de votre système.

Alarmes matérielles

Le tableau ci-dessous contient la liste des événements matériels enregistrés dans des journaux par le programme de surveillance quand une alarme est déclenchée :

Identifiant	Description	Action du programme de surveillance
0	Surchauffe – Logiciel	Arrête l'utilisation du processeur de la plateforme DRAGEN Bio-IT pour permettre un refroidissement à la température minimale d'exécution du logiciel.
1	Surchauffe – Matériel	Erreur fatale. Interrompt le logiciel DRAGEN. Un redémarrage du système est nécessaire.
2	Surchauffe – Module SPD	Enregistrée comme erreur non fatale.
3	Surchauffe – Module SODIMM	Enregistrée comme erreur non fatale.

Identifiant	Description	Action du programme de surveillance
4	Alimentation 0	Erreur fatale. Interrompt le logiciel DRAGEN. Un redémarrage du système est nécessaire.
5	Alimentation 1	Erreur fatale. Interrompt le logiciel DRAGEN. Un redémarrage du système est nécessaire.
6	Alimentation – Processeur de la plateforme DRAGEN Bio-IT	Enregistrée comme erreur non fatale.
7	Ventilateur 0	Enregistrée comme erreur non fatale.
8	Ventilateur 1	Enregistrée comme erreur non fatale.
9	SE5338	Erreur fatale. Interrompt le logiciel DRAGEN. Un redémarrage du système est nécessaire.
10 à 30	Non définies (Réservées)	Erreur fatale. Interrompt le logiciel DRAGEN. Un redémarrage du système est nécessaire.

Les alarmes d'erreurs fatales empêchent le logiciel hôte de la plateforme DRAGEN de s'exécuter et requièrent un redémarrage du système. Quand une alarme de surchauffe relative au logiciel est déclenchée, le programme de surveillance cherche et arrête tout processus de la plateforme DRAGEN en cours d'exécution. Le programme de surveillance continue d'arrêter tout nouveau processus de la plateforme DRAGEN tant que la température n'a pas diminué jusqu'au seuil minimal et que le matériel n'a pas désactivé l'alarme d'état de la puce. Quand l'alarme de surchauffe relative au logiciel est désactivée, les tâches de la plateforme DRAGEN peuvent reprendre leur exécution.

Communiquez les renseignements contenus dans les fichiers de journaux à l'assistance technique d'Illumina si une de ces alarmes se déclenche sur votre système.

Compression et décompression avec accélération matérielle

La compression au format gzip est omniprésente en bioinformatique. Les fichiers FASTQ sont souvent compressés en format gzip. Le format BAM est une version spécialisée du format gzip. C'est pourquoi le processeur de la plateforme DRAGEN Bio-IT prend en charge l'accélération matérielle de la compression et de la décompression des données en format gzip. Si vos fichiers d'entrée sont compressés en format gzip, la plateforme le détecte et les décompresse automatiquement. Si votre fichier de sortie est au format BAM, les fichiers sont automatiquement compressés.

La plateforme DRAGEN offre des utilitaires autonomes de ligne de commande qui vous permettent de compresser et de décompresser d'autres fichiers. Ces utilitaires sont similaires aux commandes gzip et gunzip de Linux, mais se nomment *dzip* et *dunzip* (pour dragen zip et dragen unzip). Ces deux utilitaires peuvent prendre en entrée un seul fichier et produisent un seul fichier de sortie auquel l'extension .gz est retirée ou ajoutée, selon le cas. Par exemple :

```
dzip file1      # produit le fichier de sortie file1.gz
dunzip file2.gz # produit le fichier de sortie file2
```

Actuellement, les commandes *dzip* et *dunzip* ont les limites et les différences suivantes par rapport aux commandes gzip et gunzip :

- Chaque invocation de ces outils ne peut accepter qu'un seul fichier. Les noms de fichier supplémentaires (y compris ceux produits par l'utilisation d'un caractère de remplacement « * ») sont ignorés.
- Elles ne peuvent être exécutées en même temps par le logiciel hôte de la plateforme DRAGEN.
- Elles ne prennent pas en charge les options de ligne de commande de gzip et gunzip (p. ex., --recursive, --fast, --best, --stdout).

Rapport sur l'utilisation

Au cours de l'installation, un programme démon (dragen_licd) est créé (ou arrêté puis redémarré). Ce processus d'arrière-plan s'active automatiquement à la fin de chaque jour pour téléverser des renseignements sur l'utilisation du logiciel hôte de la plateforme DRAGEN vers un serveur d'Illumina. Ces renseignements comprennent la date, la durée, la taille (en nombre de bases) et l'état de chaque analyse, ainsi que la version du logiciel utilisée.

La communication avec le serveur d'Illumina est sécurisée par chiffrement. S'il y a une erreur de communication, le programme démon essaie de nouveau jusqu'au matin suivant. Si le téléversement continue d'échouer, la communication est tentée de nouveau la nuit suivante jusqu'à ce qu'elle réussisse. Ceci signifie que pendant les heures de travail, les ressources du système restent entièrement disponibles et ne sont pas ralenties par cette activité d'arrière-plan.

Pour vérifier la licence actuellement utilisée, servez-vous de la commande *dragen_lic* .

Pour générer un rapport sur l'utilisation, votre serveur doit satisfaire les exigences suivantes :

- 256 Go de mémoire vive et disque dur de 2 To pour la couverture germinale 300x uniéchantillon.
- Couverture d'analyse d'échantillons de tumeur et de référence.
- 6 To et 512 Go de mémoire vive.

Le rapport sur l'utilisation ne contient pas les renseignements suivants :

- Le nombre de fichiers en entrée de la fonction de génotypage.
- Les fichiers gVCF de GATK en entrée de la fonction de génotypage de fichiers gVCF.
- Le mélange de fichiers gVCF de différentes fonctions de définition, comme les fonctions de définition combinée et de génotypage de fichiers gVCF.

Chapitre 8 Dépannage

Si le système DRAGEN ne semble pas répondre, faites ce qui suit :

1 Pour déterminer si le système DRAGEN est bloqué, suivez les instructions de la section *Comment déterminer si le système est bloqué*.

2 Recueillez les renseignements de diagnostic après un blocage ou un plantage comme décrit dans la section *Envoyer les renseignements de diagnostic à l'assistance d'Illumina*.

3 Après avoir recueilli tous les renseignements, réinitialisez votre système si nécessaire comme décrit dans la section *Réinitialiser votre système après un plantage ou un blocage*.

Comment déterminer si le système est bloqué

Le système DRAGEN a un programme de surveillance qui vérifie si le système est bloqué. Si une analyse semble prendre plus longtemps qu'elle ne devrait, le programme de surveillance pourrait ne pas avoir détecté le blocage. Voici quelques mesures à tenter :

- Exécutez la commande *top* pour déterminer le processus actif de la plateforme DRAGEN. S'il n'y a pas de problème avec votre analyse, elle devrait utiliser plus de 100 % de la puissance de l'unité centrale. Si elle en utilise 100 % ou moins, votre système pourrait être bloqué.
- Exécutez la commande *du -s* dans le répertoire du fichier de sortie BAM/SAM. Lors d'une analyse normale, la taille de ce répertoire devrait croître en raison des données de sortie intermédiaires (lorsque le tri est activé) ou des données BAM/SAM.

Envoyer les renseignements de diagnostic à l'assistance d'Illumina

Illumina aimerait recevoir vos commentaires à propos du système DRAGEN, y compris les rapports sur une défaillance du système.En cas de plantage, de blocage ou d'une défaillance du programme de surveillance, exécutez la commande *sosreport* pour recueillir les renseignements de diagnostic et de configuration comme suit :

```
sudo sosreport --batch --tmp-dir /staging/tmp
```

L'exécution de cette commande prend plusieurs minutes. Elle indique l'emplacement où elle a enregistré les renseignements de diagnostic dans le répertoire /staging/tmp.Veuillez inclure le rapport lorsque vous envoyez une demande d'assistance à l'assistance technique d'Illumina.

Réinitialiser votre système après un plantage ou un blocage

Si le système DRAGEN plante ou est bloqué, l'utilitaire *dragen_reset* doit être lancé pour réinitialiser le matériel et le logiciel. Cet utilitaire est lancé automatiquement par le logiciel hôte chaque fois qu'il détecte une condition inattendue. Dans ce cas, le logiciel hôte affiche le message suivant :

```
Running dragen_reset to reset DRAGEN Bio-IT processor and software
```

Si le logiciel est bloqué, veuillez recueillir les renseignements de diagnostic comme décrit dans la sous-section *Envoyer les renseignements de diagnostic à l'assistance d'Illumina* à la page 161, puis lancez manuellement la commande *dragen_reset* comme suit :

```
/opt/edico/bin/dragen_reset
```

Toute exécution de la commande *dragen_reset* nécessitera un rechargement du génome de référence dans la carte DRAGEN. Le logiciel hôte recharge automatiquement la référence lors de la prochaine exécution.

Annexe A Options de ligne de commande

Options générales du logiciel

Les options suivantes se trouvent dans la section par défaut du fichier de configuration. La section par défaut n'est associée à aucun nom de section (p. ex., [Aligner]). La section par défaut se trouve au début du fichier de configuration. Veuillez prendre note que certains champs obligatoires doivent être spécifiés dans la ligne de commande et ne sont pas présents dans les fichiers de configuration.

Nom	Description	Équivalent dans la ligne de commande	Valeurs
alt-aware	Active un traitement spécial des contigs ALT, si un fichier LiftOver ALT a été utilisé dans la table de hachage. Option activée par défaut si une référence a été créée grâce à un fichier LiftOver.	--alt-aware	true/false
annotation-file	Fichier d'annotations de transcrits (ARN).	--annotation-file, -a	
append-read-index-to-name	Par défaut, la plateforme DRAGEN donne le même nom aux deux extrémités d'une paire. Lorsque sa valeur est mise à « true » (activée), la plateforme DRAGEN ajoute « /1 » et « /2 » aux deux extrémités.		true/false
bam-input	Le fichier BAM aligné d'entrée pour la fonction de définition de variants de la plateforme DRAGEN.	-b, --bam-input	
bcl-conversion-only	Effectue une conversion du format BCL d'Illumina au format FASTQ.	--bcl-conversion-only	
bcl-input-directory	Le répertoire BCL d'entrée pour la conversion au format BCL.	--bcl-input-directory	
sample-sheet	Le chemin vers le fichier SampleSheet.csv pour l'entrée BCL. L'emplacement par défaut est le répertoire racine BCL.	--sample-sheet	
bcl-use-hw	Mettez la valeur « false » (désactivée) pour empêcher l'utilisation de l'accélération FPGA de la plateforme DRAGEN durant la conversion au format BCL. La valeur par défaut est « true » (activée).	--bcl-use-hw	true/false
build-hash-table	Génère une référence/table de hachage.	--build-hash-table	true/false
cram-input	Produit un fichier CRAM pour la fonction de définition de variants DRAGEN.	--cram-input	
dbsnp	Le chemin vers le fichier VCF (or .vcf.gz) de la base de données d'annotation des variants.	--dbsnp	
enable-auto-multifile	Importe les segments subséquents des fichiers *_001.{dbam,fastq}.	--enable-auto-multifile	true/false
enable-bam-indexing	Permet la génération d'un fichier d'indexation BAI.	--enable-bam-indexing	true/false
enable-cnv	Active le traitement de variants du nombre de copies (VNC).	--enable-cnv	true/false
enable-duplicate-marking	Permet de marquer les enregistrements d'alignements répétés produits.	--enable-duplicate-marking	true/false

Nom	Description	Équivalent dans la ligne de commande	Valeurs
enable-map-align-output	Permet l'enregistrement de la sortie de l'étape de la cartographie et de l'alignement. La valeur par défaut est « true » (activée) lorsque seulement la cartographie et l'alignement sont effectués. La valeur par défaut est « false » (désactivée) lorsque la fonction de définition de variants est utilisée.	--enable-map-align-output	true/false
enable-methylation-calling	Indique s'il faut automatiquement ajouter des étiquettes de méthylation et produire un seul fichier BAM pour les protocoles de méthylation.		true/false
enable-rna	Permet de traiter des données de séquençage de l'ARN.	--enable-rna	true/false
enable-rna-quantification	Permet d'activer ou de désactiver la quantification de l'ARN. Lorsque la valeur « true » (activée) est sélectionnée, la valeur de l'option enable-rna doit également être mise à « true » (activée).	--enable-rna-quantification	true/false
enable-sampling	Détecte automatiquement les paramètres des lectures appariées en analysant un échantillon avec les fonctions de cartographie et d'alignement.		true/false
enable-sort	Active le tri après la cartographie et l'alignement.		true/false
enable-variant-caller	Active la fonction de définition de variants.	--enable-variant-caller	true/false
enable-vcf-compression	Permet la compression des fichiers VCF de sortie. La valeur par défaut est « true » (activée).		true/false
fastq-file1	Le fichier FASTQ d'entrée du pipeline de la plateforme DRAGEN (peut être compressé en format gzip).	-1, --fastq-file1	
fastq-file2	Le deuxième fichier FASTQ d'entrée avec des lectures appariées.	-2, --fastq-file2	
fastq-list	Le fichier CSV contenant une liste de fichiers FASTQ à traiter.	--fastq-list	
fastq-list-all-samples	Permet d'activer ou de désactiver le traitement de tous les échantillons du même coup, sans égard à la valeur RGSM.	--fastq-list-all-samples	true/false
fastq-n-quality	La qualité de la définition de bases à produire pour les bases N. Ajoutée automatique à l'option fastq-n-quality pour tous les fichiers de sortie N.	--fastq-n-quality	0 à 255
fastq-offset	La valeur de décalage de qualité de fichier FASTQ.	--fastq-offset	33 ou 64
filter-flags-from-output	Filtre les alignements produits selon les valeurs inscrites dans le champ Flags (étiquettes). Les valeurs hexadécimales et décimales sont acceptées.	--filter-flags-from-output	
first-tile-only	Permet de convertir uniquement la première plaque de chaque ligne durant la conversion au format BCL.	--first-tile-only	
force	Force le remplacement du fichier de sortie existant.	-f	
force-load-reference	Force le chargement de la référence et des tables de hachage avant de lancer le pipeline de la plateforme DRAGEN.	-l	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
generate-md-tags	Permet de générer ou non des étiquettes MD avec les enregistrements d'alignements de sortie. La valeur par défaut est « false » (désactivée).	--generate-md-tags	true/false
generate-sa-tags	Permet de générer ou non des étiquettes SA:Z pour les enregistrements comprenant des alignements chimériques ou supplémentaires.	--generate-sa-tags	true/false
generate-zs-tags	Permet de générer ou non des étiquettes ZS pour les enregistrements d'alignements de sortie. La valeur par défaut est « false » (désactivée).	--generate-sz-tags	true/false
ht-alt-liftover	Fichier LiftOver en format SAM des contigs alternatifs de référence.	--ht-alt-liftover	
ht-alt-aware-validate	Désactive l'exigence d'un fichier LiftOver lors de la construction d'une table de hachage à partir d'une référence qui contient des contigs ALT.	--ht-alt-aware-validate	true/false
ht-build-rna-hashtable	Permet de générer une table de hachage de l'ARN. La valeur par défaut est « false » (désactivée).	--ht-build-rna-hashtable	true/false
ht-cost-coeff-seed-freq	Le coefficient des coût pour la fréquence de point d'ancrage étendu.	--ht-cost-coeff-seed-freq	
ht-cost-coeff-seed-len	Le coefficient des coût pour la longueur de point d'ancrage étendu.	--ht-cost-coeff-seed-len	
ht-cost-penalty-incr	La coût supplémentaire pour chaque étape d'incrémentation d'extension de point d'ancrage.	--ht-cost-penalty-incr	
ht-cost-penalty	Le coût supplémentaire pour l'extension d'un point d'ancrage par un certain nombre de bases.	--ht-cost-penalty	
ht-decoys	Spécifie le chemin d'accès vers un fichier de leurres.	--ht-decoys	
ht-max-dec-factor	Le facteur maximal d'élimination pour la réduction du nombre de points d'ancrage.	--ht-max-dec-factor	
ht-max-ext-incr	Le nombre maximal de bases d'une étape d'extension de point d'ancrage.	--ht-max-ext-incr	
ht-max-ext-seed-len	La longueur maximale de point d'ancrage étendu.	--ht-max-ext-seed-len	
ht-max-seed-freq	La fréquence maximale pour une correspondance de point d'ancrage après des tentatives d'extension.	--ht-max-seed-freq	1–256
ht-max-table-chunks	Le nombre maximal de blocs de table de hachage d'environ 1 Go de mémoire (simultanément).	--ht-max-table-chunks	
ht-mem-limit	La limite de mémoire (table de hachage + référence), unités octet|ko|Mo|Go.	--ht-mem-limit	
ht-methylated	Génère automatiquement des tables de hachage de référence avec conversions C->T et G->A.	--ht-methylated	true/false
ht-num-threads	Le nombre maximal de fils de travail de l'unité centrale utilisés pour construire une table de hachage.	--ht-num-threads	
ht-rand-hit-extend	Pour inclure une occurrence aléatoire avec chaque enregistrement EXTEND de cet enregistrement de fréquence.	--ht-rand-hit-extend	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
ht-rand-hit-hifreq	Pour inclure une occurrence aléatoire avec chaque enregistrement HIFREQ.	--ht-rand-hit-hifreq	
ht-ref-seed-interval	Le nombre de positions par point d'ancrage de référence.	--ht-ref-seed-interval	
ht-reference	Le fichier de référence en format .fasta à utiliser pour construire une table de hachage.	--ht-reference	
ht-seed-len	La longueur initiale du point d'ancrage à ajouter dans la table de hachage.	--ht-seed-len	
ht-size	La taille de la table de hachage, unités octet|ko|Mo|Go.	--ht-size	
ht-soft-seed-freq-cap	La limite souple de fréquence pour la réduction du nombre de points d'ancrage.	--ht-soft-seed-freq-cap	
ht-suppress-decoys	Supprime l'utilisation des fichiers de leurres lors de la construction d'une table de hachage.	--ht-suppress-decoys	
ht-target-seed-freq	La fréquence de points d'ancrage cible pour l'extension de points d'ancrage.	--ht-target-seed-freq	
input-qname-suffix-delimiter	Spécifie le délimiteur utilisé par l'option append-read-index-to-name et pour la détection des noms de paire correspondants à ceux du fichier BAM d'entrée.		/ou . ou :
interleaved	Les lectures appariées imbriquées dans un seul fichier FASTQ.	-i	
intermediate-results-dir	Le répertoire où stocker les résultats intermédiaires (p. ex., tri des partitions).		
lic-no-print	Supprime le message d'état de la licence à la fin de l'analyse.	--lic-no-print	true/false
mapq-strict-js	Propre à l'ARN. Si la valeur est de « 0 », le score de qualité MAPQ sera plus élevé, ce qui exprime la confiance que l'alignement est au moins partiellement correct. Si la valeur est de « 1 », le score de qualité MAPQ sera plus faible, ce qui exprime une ambiguïté de jonction d'épissage.	--mapq-strict-js	0/1
methylation-generate-cytosine-report	Génère un rapport de méthylation de la cytosine du génome entier.	--methylation-generate-cytosine-report	true/false
methylation-generate-mbias-report	Génère un rapport de biais de méthylation par cycle de séquenceur.		true/false
methylation-match-bismark	Si la valeur est mise à « true » (activée), les étiquettes de Bismark sont exactement les mêmes, y compris les bogues.	--methylation-match-bismark	true/false
methylation-protocol	Le protocole de librairie pour l'analyse de la méthylation.	--methylation-protocol	none / directional / nondirectional / directional-complement
num-threads	Le nombre de fils de processeur à utiliser.	-n, --num-threads	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
output-directory	Le répertoire de sortie.	--output-directory	
output-file-prefix	Le préfixe du nom de fichier de sortie à utiliser pour tous les fichiers générés par le pipeline.	--output-file-prefix	
output-format	Le format du fichier de sortie de l'étape de cartographie et d'alignement. Les valeurs valides sont BAM (par défaut), SAM, ou DBAM (un format binaire exclusif).	--output-format	BAM/SAM/DBAM
pair-by-name	Spécifie si l'ordre des enregistrements d'entrée BAM doit être modifié afin que les couples appariés soient traités ensemble.		
pair-suffix-delimiter	Modifie le caractère de délimitation des suffixes.	--pair-suffix-delimiter	/. :
preserve-bqsr-tags	Indique s'il faut conserver ou non les étiquettes BI et BD du fichier BAM d'entrée. Remarque : Cela peut créer des problèmes dans le cas d'une base coupée absente de l'alignement.		true/false
preserve-map-align-order	Produit un fichier de sortie qui conserve l'ordre original des lectures du fichier d'entrée.		true/false
qc-coverage-region-1	Le premier fichier BED dont la couverture fait l'objet d'un rapport.	--qc-coverage-region-1	
qc-coverage-region-2	Le deuxième fichier BED dont la couverture fait l'objet d'un rapport.	--qc-coverage-region-2	
qc-coverage-region-3	Le troisième fichier BED dont la couverture fait l'objet d'un rapport.	--qc-coverage-region-3	
qc-coverage-reports-1	Les types de rapports demandés pour l'option qc-coverage-region-1.	--qc-coverage-reports-1	full_res/cov_report
qc-coverage-reports-2	Les types de rapports demandés pour l'option qc-coverage-region-2.	--qc-coverage-reports-2	full_res/cov_report
qc-coverage-reports-3	Les types de rapports demandés pour l'option qc-coverage-region-3.	--qc-coverage-reports-3	full_res/cov_report
ref-dir	Le répertoire contenant la table de hachage de référence. Cette référence est automatiquement chargée dans la carte DRAGEN, si elle ne s'y trouve pas déjà.	-r, --ref-dir	
ref-sequence-filter	Pour ne produire que les lectures qui correspondent à cette séquence de référence.	--ref-sequence-filter	
remove-duplicates	Si la valeur est « true » (activée), les enregistrements d'alignements répétés sont supprimés au lieu d'être simplement signalés à l'aide d'une étiquette.		true/false
RGCN	Le nom du centre de séquençage d'un groupe de lectures.	--RGCN	
RGCN-tumor	Le nom du centre de séquençage d'un groupe de lectures pour une entrée de tumeur.	--RGCN-tumor	
RGDS	La description du groupe de lectures.	--RGDS	
RGDS-tumor	La description du groupe de lectures pour une entrée de tumeur.	--RGDS-tumor	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
RGDT	La date d'analyse du groupe de lectures.	--RGDT	
RGDT-tumor	La date d'analyse du groupe de lectures pour une entrée de tumeur.	--RGDT-tumor	
RGID	L'identifiant du groupe de lectures.	--RGID	
RGID-tumor	L'identifiant du groupe de lectures pour une entrée de tumeur.	--RGID-tumor	
RGLB	La librairie du groupe de lectures.	--RGLB	
RGLB-tumor	La librairie du groupe de lectures pour une entrée de tumeur.	--RGLB-tumor	
RGPI	La taille prévue des inserts du groupe de lectures.	--RGPI	
RGPI-tumor	La taille prévue des inserts du groupe de lectures pour une entrée de tumeur.	--RGPI-tumor	
RGPL	La technologie de séquençage du groupe de lectures.	--RGPL	
RGPL-tumor	La technologie de séquençage du groupe de lectures pour une entrée de tumeur.	--RGPL-tumor	
RGPU	L'unité de plateforme du groupe de lectures.	--RGPU	
RGPU-tumor	L'unité de plateforme du groupe de lectures pour une entrée de tumeur.	--RGPU-tumor	
RGSM	Le nom de l'échantillon du groupe de lectures.	--RGSM	
RGSM-tumor	Le nom de l'échantillon du groupe de lectures pour une entrée de tumeur.	--RGSM-tumor	
rna-ann-sj-min-len	Retire les jonctions d'épissage dont la longueur est inférieure à cette valeur lors de la génération de ces jonctions à partir d'un fichier d'annotations (GTF/GFF/SJ.out.tab).		
rna-gf-input-file	Un fichier .Chimeric.out.junction déjà généré. Lorsque ce fichier est fourni, le module de fusion de gènes de la plateforme DRAGEN est exécuté en mode autonome.	--rna-gf-input-file	
rna-mapq-unique	Aux fins de compatibilité avec Cufflinks, activez ce paramètre avec une valeur autre que zéro. Les fonctions de cartographie uniques ont un score de qualité MAPQ configuré pour cette valeur. Les fonctions de cartographie multiples ont un score de qualité de int(-10*log10(1 – 1 /NH)).		« 0 », « 1 » à « 255 »
rna-quantification-fld-max	Spécifie la distribution de la taille des inserts de la librairie de séquençage de l'ARN pour les analyses à paire de bases uniques. La valeur par défaut est « 250 +- 25 ».	--rna-quantification-fld-max	La valeur maximale est « 1000 ».
rna-quantification-fld-mean		--rna-quantification-fld-mean	
rna-quantification-fld-sd		--rna-quantification-fld-sd	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
rna-quantification-library-type	Spécifie le type de librairie de séquençage de l'ARN. La valeur par défaut est « A », autodétection.	--rna-quantification-library-type	IU, ISR, ISF, U, SR, SF ou A.
sample-size	Nombre de lectures de l'échantillon lorsque la valeur de l'option enable-sampling est « true » (activée).		
sample-sex	Le sexe de l'échantillon.	--sample-sex	
strict-mode	Interrompt le processus si des fichiers sont manquants.	--strict-mode	
strip-input-qname-suffixes	Indique si les suffixes d'index de lectures (p. ex., « /1 » et « /2 ») sont retirés des champs QNAME en entrée.		true/false
tumor-bam-input	Le fichier BAM aligné pour la fonction de définition de variants de la plateforme DRAGEN en mode somatique.	--tumor-bam-input	
tumor-cram-input	Le fichier CRAM aligné pour la fonction de définition de variants de la plateforme DRAGEN en mode somatique.	--tumor-cram-input	
tumor-fastq-list	Le fichier CSV contenant une liste de fichiers FASTQ pour les fonctions de cartographie, d'alignement et de définition de variants somatiques.	--tumor-fastq-list	
tumor-fastq-list-sample-id	L'identifiant d'échantillon pour la liste de fichiers FASTQ spécifiée par l'option tumor-fastq-list.	--tumor-fastq-list-sample-id	
tumor-fastq1	Le fichier FASTQ destiné au pipeline de la plateforme DRAGEN et utilisant la fonction de définition de variants en mode somatique (peut être compressé en format gzip).	--tumor-fastq1	
tumor-fastq2	Le deuxième fichier FASTQ, comprenant des lectures appariées aux lectures du fichier tumor-fastq1 et destiné au pipeline de la plateforme DRAGEN, utilisant la fonction de définition de variants en mode somatique (peut être compressé en format gzip).	--tumor-fastq2	
umi-correction-scheme	La méthodologie utilisée pour corriger les erreurs de séquençage des identifiants moléculaires uniques (IMU).	--umi-correction-scheme	Lookup/random/none
umi-enable	Permet le traitement des lectures basé sur les identifiants moléculaires uniques (IMU).	--umi-enable	true/false
umi-enable-duplex-merging	Indique si la fusion de familles est permise dans les cas d'identifiants moléculaires uniques (IMU) fractionnés et complémentaires.	--umi-enable-duplex-merging	true/false
umi-min-supporting-reads	Le nombre de lectures d'entrée dont l'identifiant moléculaire unique (IMU) et la position correspondante sont nécessaires pour générer une lecture de consensus.	--umi-min-supporting-reads	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
umi-random-merge-factor	Aux fins de corrections aléatoires d'identifiants moléculaires uniques (IMU), le rapport du dénombrement de fragments nécessaire pour corriger et fusionner deux familles ayant des identifiants moléculaires uniques (IMU) semblables.	--umi-random-merge factor	
verbose	Permet à la plateforme DRAGEN de produire une sortie détaillée.	-v	
version	Affiche la version et quitte.	-V	

Options de la fonction de cartographie

Les options suivantes se trouvent dans la section [Mapper] du fichier de configuration. Pour obtenir plus de renseignements sur ces options, consultez la section *Cartographie de l'ADN* à la page 16.

Nom	Description	Équivalent dans la ligne de commande	Valeurs
ann-sj-max-indel	La longueur maximale d'indel que l'on prévoit trouver près d'une jonction d'épissage annotée.	--Mapper.ann-sj-max-indel	0 à 63
edit-chain-limit	Lorsque la valeur de l'option edit-mode est « 1 » ou « 2 » : la longueur maximale d'une chaîne de points d'ancrage dans une lecture pour qu'elle soit admissible à la modification de point d'ancrage.	--Mapper.edit-chain-limit	edit-chain-limit >= 0
edit-mode	0 = Aucune modification, 1 = Test de la longueur de la chaîne, 2 = Test de la longueur de la chaîne appariée, 3 = Modification de tous les points d'ancrage standards.	--Mapper.edit-mode	0 à 3
edit-read-len	Lorsque la valeur de l'option edit-mode est « 1 » ou « 2 » : la longueur de la lecture pour laquelle est permise la modification des points d'ancrage selon l'option edit-seed-num.	--Mapper.edit-read-len	edit-read-len > 0
edit-seed-num	Lorsque la valeur de edit-mode est « 1 » ou « 2 » : le nombre demandé de points d'ancrage par lecture qui peuvent être modifiés.	--Mapper.edit-seed-num	edit-seed-num >= 0
enable-map-align	Permet l'utilisation de fichiers BAM en entrée de la fonction de cartographie et d'alignement.	--enable-map-align	true/false
map-orientations	0 = ordre normal, 1 = ordre normal seulement, 2 = ordre de complément inverse seulement (les extrémités appariées requièrent l'ordre normal).	--Mapper.map-orientations	0 à 2
max-intron-bases	La longueur maximale d'intron signalé.	--Mapper.max-intron-bases	
min-intron-bases	La longueur minimale de délétion de référence signalée comme un intron.	--Mapper.min-intron-bases	
seed-density	La densité de points d'ancrage exigée des lectures demandées dans la table de hachage.	--Mapper.seed-density	0 > seed-density > 1

Options de la fonction d'alignement

Les options suivantes se trouvent dans la section [Aligner] du fichier de configuration. Pour obtenir plus de renseignements, consultez la section *Alignement de l'ADN* à la page 19

Nom	Description	Équivalent dans la ligne de commande	Valeurs
aln-min-score	(signé) Le score minimal d'alignement à signaler; référence pour le score MAPQ. Lorsque des alignements locaux (global = 0) sont utilisés, l'option aln-min-score est calculé par le logiciel hôte comme ayant une valeur « 22 * match-score ». Lorsque des alignements globaux sont utilisés (global = 1), la valeur de l'option aln-min-score est « -1 000 000 ». Le calcul du logiciel hôte peut être remplacé en définissant la valeur de l'option aln-min-score dans le fichier de configuration.	--Aligner.aln-min-score	–2 147 483 648 à 2 147 483 647
dedup-min-qual	La qualité de base minimale pour le calcul de l'indicateur de qualité des lectures pour la déduplication.	--Aligner.dedup-min-qual	0–63
en-alt-hap-aln	Permet aux alignements chimériques d'être produits, comme alignements supplémentaires.	--Aligner.en-alt-hap-aln	0–1
en-chimeric-aln	Permet aux alignements chimériques d'être produits, comme alignements supplémentaires.	--Aligner.en-chimeric-aln	0–1
gap-ext-pen	Le score de pénalité pour un écart d'extension.	--Aligner.gap-ext-pen	0–15
gap-open-pen	Le score de pénalité pour l'intégration d'un écart (insertion ou délétion).	gap-open-pen	0–127
global	Indique si un alignement est global (Needleman-Wunsch) plutôt que local (Smith-Waterman).	--Aligner.global	0–1
hard-clips	Les marques de découpage : [0] primary (primaire), [1] supplementary (supplémentaire), [2] secondary (secondaire).	--Aligner.hard-clips	3 bits
map-orientations	Limite les orientations à « forward-only », « reverse-complement only » ou « any alignments ».	--Aligner.map-orientations	0 (n'importe quel alignement) 1 (dans l'ordre normal seulement) 2 (complément inverse seulement)
mapq-max	Le plafond du score rapporté de la qualité de la cartographie (MAPQ).	--Aligner.mapq-max	0 à 255
match-n-score	(signé) L'incrémentation du score quand il y a correspondance avec un code IUB « N » pour un nucléotide de référence.	--Aligner.match-n-score	-16–15

Nom	Description	Équivalent dans la ligne de commande	Valeurs
match-score	L'incrémentation du score quand il y a correspondance avec un nucléotide de référence.	--Aligner.match-score	Lorsque la valeur de l'option global = 0, la valeur de match-score est supérieure à 0. Lorsque la valeur de l'option global = 1, la valeur de match-score est supérieure ou égale à 0.
max-rescues	Le nombre maximal d'alignements récupérés par paire de lectures. La valeur par défaut est « 10 ».	--max-rescues	0–1023
min-score-coeff	L'ajustement effectué à la valeur de l'option aln-min-score pour chaque lecture.	--Aligner. min-score-coeff	-64–63,999
mismatch-pen	Le score de pénalité pour un mésappariement.	--Aligner.mismatch-pen	0–63
no-unclip-score	Lorsque l'option no-unclip-score a une valeur de « 1 », toute contribution non découpée supplémentaire à un alignement est retirée du score de cet alignement avant que le traitement ne soit effectué.	--Aligner.no-unclip-score	0-1
no-unpaired	Définit si seuls les alignements bien appariés doivent être rapportés pour les lectures appariées.	--Aligner. no-unpaired	0–1
pe-max-penalty	La pénalité maximale au score d'appariement, pour les extrémités non appariées ou distantes.	--Aligner.pe-max-penalty	0–255
pe-orientation	L'orientation attendue des lectures appariées : 0 = FR, 1 = RF, 2 = FF.	--Aligner.pe-orientation	0, 1, 2
rescue-sigmas	Les écarts de la longueur de lecture moyenne utilisée comme portée d'analyse des récupérations. La valeur par défaut est « 2,5 ».	--Aligner.rescue-sigmas	
sec-aligns	Le nombre maximal d'alignements secondaires (sous-optimaux) à rapporter par lecture.	--Aligner.sec-aligns	0–30
sec-aligns-hard	Permet de forcer les lectures non cartographiées lorsque les alignements secondaires ne sont pas tous produits.	--Aligner.sec-aligns-hard	0–1
sec-phred-delta	Seuls les alignements secondaires dont les probabilités se situent dans la même échelle de valeurs Phred que l'alignement primaire sont signalés.	--Aligner.sec-phred-delta	0–255
sec-score-delta	Les alignements secondaires sont autorisés si leur score d'appariement n'est pas plus faible que cette limite (par rapport à l'alignement primaire).	--Aligner. sec-score-delta	
supp-aligns	Le nombre maximal d'alignements supplémentaires (chimériques) à rapporter par lecture.	--Aligner.supp-aligns	0–30

Nom	Description	Équivalent dans la ligne de commande	Valeurs
supp-as-sec	Permet de rapporter les alignements supplémentaires avec un indicateur d'alignement secondaire.	--Aligner.supp-as-sec	0–1
supp-min-score-adj	L'augmentation du score d'alignement minimal pour obtenir des alignements supplémentaires. Ce score est calculé par le logiciel hôte comme « 8 * match-score » pour l'ADN et sa valeur par défaut est « 0 » pour l'ARN.	--Aligner. supp-min-score-adj	
unclip-score	La bonification du score pour atteindre chaque extrémité de la lecture.	--Aligner.unclip-score	0–127
unpaired-pen	La pénalité pour les alignements non appariés dans l'échelle Phred.	--Aligner.unpaired-pen	0–255

Si vous avez désactivé la détection automatique des statistiques de longueur d'insertion à l'aide de l'option *--enable-sampling*, vous devez remplacer toutes les options suivantes pour spécifier les statistiques. Pour plus de renseignements, consultez la section *Détection de la taille moyenne des inserts* à la page 23. Les options suivantes se trouvent dans la section [Aligner] du fichier de configuration.

Option	Description	Équivalent dans la ligne de commande	Valeurs
pe-stat-mean-insert	La longueur moyenne du modèle.		0–65535
pe-stat-mean-read-len	La longueur moyenne de la lecture.		0–65535
pe-stat-quartiles-insert	Le trio de nombres délimités par des virgules qui spécifie les longueurs de modèle appartenant aux 25^{e}, 50^{e}, and 75^{e} percentile.		0–65535
pe-stat-stddev-insert	L'écart-type de la répartition des longueurs de modèle.		0–65535

Options de la fonction de définition de variants

Les options suivantes se trouvent dans la section « Variant Caller » du fichier de configuration. Pour plus de renseignements, consultez la section *Options de la fonction de définition de variants* à la page 31.

Nom	Description	Équivalent dans la ligne de commande	Valeurs
cnv-blacklist-bed	Spécifie un fichier BED indiquant des intervalles à exclure du fichier de sortie final.	--cnv-blacklist-bed	
cnv-cbs-alpha	Le niveau d'importance à partir duquel le test accepte les modifications. La valeur par défaut est « 0.01 ».	cnv-cbs-alpha	
cnv-cbs-eta	Le taux d'erreur de type I des limites séquentielles pour un arrêt précoce lorsque la méthode de permutation est utilisée. La valeur par défaut est « 0,05 ».	--cnv-cbs-eta	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
cnv-cbs-kmax	La largeur maximale du plus petit segment destiné à la permutation. La valeur par défaut est « 25 ».	--cnv-cbs-kmax	
cnv-cbs-min-width	Le nombre minimal de marqueurs pour un segment modifié. La valeur par défaut est « 2 ».	--cnv-cbs-min-width	
cnv-cbs-nmin	La longueur minimale de données pour une approximation statistique maximale. La valeur par défaut est « 200 ».	--cnv-cbs-nmin	
cnv-cbs-nperm	Le nombre de permutations utilisées dans le calcul de la valeur p. La valeur par défaut est « 10 000 ».	--cnv-cbs-nperm	
cnv-cbs-trim	La proportion des données à retrancher aux fins de calculs de la variance. La valeur par défaut est « 0,025 ».	--cnv-cbs-trim	
cnv-counts-method	La méthode de dénombrement pour un alignement à dénombrer dans un compartiment cible. La méthode par défaut est « Overlap » (chevauchement) pour une approche par panel de référence. Dans le cas de la normalisation automatique, la valeur par défaut est « Start » (départ).	--cnv-counts-method	midpoint/start/overlap
cnv-enable-gcbias-correction	Permet d'activer ou de désactiver la correction de biais GC.	--cnv-enable-gcbias-correction	true/false
cnv-enable-gcbias-smoothing	Permet d'activer ou de désactiver la correction de biais GC simplifiée au sein de compartiments GC ayant un noyau exponentiel. La valeur par défaut est « true » (activée).	--cnv-enable-gcbias-smoothing	true/false
cnv-enable-plots	Permet d'activer ou de désactiver la génération de tracés. La valeur par défaut est « false » (désactivée).	--cnv-enable-plots	true/false
cnv-enable-ref-calls	Lorsque la valeur est « true » (activée), les définitions de copies neutres (REF) sont comprises dans le fichier VCF de sortie.	--cnv-enable-ref-calls	true/false
cnv-enable-self-normalization	Permet d'activer ou de désactiver la normalisation automatique.	--cnv-enable-self-normalization	true/false
cnv-enable-tracks	Permet d'activer ou de désactiver la génération de fichiers de suivi pouvant être importés dans l'outil Integrative Genome Viewer (IGV). La valeur par défaut est « true » (activée).	--cnv-enable-tracks	true/false
cnv-extreme-percentile	La valeur seuil de percentile médian utilisée pour filtrer les échantillons. La valeur par défaut est « 2,5 ».	--cnv-extreme-percentile	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
cnv-filter-bin-support-ratio	Filtre un événement candidat si l'étendue des compartiments connexes est inférieure au rapport spécifié en concordance avec la longueur globale de l'événement. Le rapport par défaut est « 0,2 » (20 %).	--cnv-filter-bin-support-ratio	
cnv-filter-copy-ratio	Le rapport minimal du nombre de copies, centré autour de « 1,0 », à partir duquel un événement est marqué « PASS » (réussite) dans le fichier de sortie VCF. La valeur par défaut est « 0,2 ».	--cnv-filter-copy-ratio	
cnv-filter-de-novo-quality	Le seuil, sur l'échelle Phred, de définition d'un événement de novo chez le proposant.	--cnv-filter-de-novo-quality	
cnv-filter-length	Longueur maximale d'un événement (dans les bases) pour laquelle un événement d'un rapport est marqué « PASS » (réussite) dans le fichier VCF de sortie. La valeur par défaut est « 10 000 ».	--cnv-filter-length	
cnv-filter-qual	La valeur QUAL pour laquelle un événement signalé est marqué « PASS » (réussite) dans le fichier VCF de sortie. La valeur par défaut est « 10 ».	--cnv-filter-qual	
cnv-fpop-penalty	La paramètre de pénalité pour la détection des modifications. La valeur par défaut est « 0,03 ».	--cnv-fpop-penalty	
cnv-input	L'échantillon d'entrée.	--cnv-input	
cnv-interval-width	La largeur de l'intervalle d'échantillonnage pour le séquençage du génome entier de variants du nombre de copies (VCN). La valeur par défaut est « 1000 ».	--cnv-wgs-interval-width	
cnv-matched-normal	Le fichier de dénombrements de cibles de l'échantillon de référence correspondant.	--cnv-matched-normal	
cnv-max-percent-zero-samples	Le seuil pour le filtrage de cibles lorsqu'il y a trop d'échantillons sans couverture. La valeur par défaut est « 5 % ».	--cnv-max-percent-zero-samples	
cnv-max-percent-zero-targets	Le seuil pour le filtrage d'échantillons lorsqu'il y a trop de cibles sans couverture. La valeur par défaut est « 2,5 % ».	--cnv-max-percent-zero-targets	
cnv-merge-distance	Spécifie le nombre minimal de paires de bases entre deux segments afin de permettre leur fusion. La valeur par défaut est « 0 », c'est-à-dire que les segments doivent être directement adjacents.	--cnv-merge-distance	
cnv-merge-threshold	La différence maximale entre les valeurs moyennes des segments à laquelle deux segments adjacents devraient être fusionnés. La valeur moyenne des segments est représentée par un rapport de copies linéaire. La valeur par défaut est « 0,2 ». La valeur 0 désactive la fusion.	--cnv-merge-threshold	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
cnv-min-mapq	Permet d'activer ou de désactiver le fractionnement de tous les intervalles de fichier BED de cibles en deux intervalles équivalents. Lorsque le fractionnement est activé, il doit l'être pour tous les échantillons (l'échantillon à analyser et le panel de référence). La valeur par défaut est « false » (désactivée).	--cnv-min-mapq	true/false
cnv-normals-file	Le fichier unique à utiliser dans le panel de référence. Cette option peut être utilisée plusieurs fois, une par fichier.	--cnv-normals-file	
cnv-normals-list	Le fichier du panel de référence.	--cnv-normals-list	
cnv-num-gc-bins	Le nombre de compartiments pour la correction de biais GC. Chaque compartiment représente un pourcentage de contenu GC. La valeur par défaut est « 25 ».	--cnv-num-gc-bins	10/20/25/50/100
cnv-ploidy	La valeur de ploïdie de référence. Cette option est utilisée pour l'estimation du nombre de copies produite dans le fichier VCF de sortie. La valeur par défaut est « 2 ».	--cnv-ploidy	
cnv-qual-length-scale	Le facteur de pondération de biais visant à ajuster les estimations QUAL pour les segments plus longs. Une option avancée qui ne devrait pas avoir à être modifiée. La valeur par défaut est « 0,9303 » (2-0,1).	--cnv-qual-length-scale	
cnv-qual-noise-scale	Le facteur de pondération de biais pour ajuster les estimations QUAL à partir de la variance de l'échantillon. Une option avancée qui ne devrait pas avoir à être modifiée. La valeur par défaut est « 1,0 ».	--cnv-qual-noise-scale	
cnv-segmentation-mode	L'algorithme de segmentation à appliquer. La valeur par défaut est « slm » ou « cbs », selon si les intervalles sont des intervalles de génomes entiers ou des intervalles de séquençage ciblé.	--cnv-segmentation-mode	cbs/slm/hslm/fpop
cnv-skip-contig-list	La liste, à valeurs séparées par des virgules, d'identifiants de contigs à sauter lors de la génération d'intervalles pour l'analyse de séquençage du génome entier. Par défaut, les contigs qui sont sautés, s'ils ne sont pas spécifiés, sont « chrM,MT,m,chrm ».	--cnv-wgs-skip-contig-list	
cnv-slm-eta	La probabilité de référence que la valeur du processus moyen change. La valeur par défaut est « 1e-5 ».	--cnv-slm-eta	
cnv-slm-fw	Le nombre minimal de points de données pour signaler un variant du nombre de copies (VNC). La valeur par défaut est « 0 ».	--cnv-slm-fw	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
cnv-slm-omega	Le paramètre d'échelle modulant la pondération relative entre la variance expérimentale/biologique. La valeur par défaut est « 0,3 ».	--cnv-slm-omega	
cnv-slm-stepeta	Le paramètre de normalisation de la distance. La valeur par défaut est « 10 000 ». Valide seulement pour la segmentation « HSLM ».	--cnv-slm-stepeta	
cnv-target-bed	Le fichier adéquatement formaté (au format BED) spécifiant l'intervalle de cibles couvert par l'échantillon. À utiliser pour les analyses de séquençage d'exome entier.	--cnv-target-bed	
--cnv-target-factor-threshold	Le pourcentage de seuil de facteur cible médian pour filtrer les cibles utilisables. La valeur par défaut est « 1 % » pour le traitement du génome entier et « 10 % » pour le traitement du séquençage ciblé.	--cnv-target-factor-threshold	
cnv-truncate-threshold	Les valeurs aberrantes extrêmes sont tronquées conformément à ce pourcentage de seuil. La valeur par défaut est « 0,1 % ».	--cnv-truncate-threshold	
cosmic	Le fichier VCF de COSMIC d'entrée.	--cosmic	
--dn-cnv-vcf	Le fichier VCF de variants structurels combiné provenant de l'étape de définition de variants du nombre de copies (VNC). En cas d'omission, les vérifications avec les VNC qui se chevauchent sont sautées.	--dn-cnv-vcf	
dn-input-vcf	Le fichier VCF de petits variants combiné à filtrer après l'étape de définition de novo.	--dn-input-vcf	
dn-output-vcf	L'emplacement pour l'écriture du fichier VFC filtré. Si l'emplacement n'est pas spécifié, le fichier VCF d'entrée est écrasé.	dn-output-vcf	
dn-sv-vcf	Le fichier VCF de variants structurels combiné provenant de l'étape de définition de variants structurels. En cas d'omission, les vérifications avec les variants structurels qui se chevauchent sont sautées.	--dn-sv-vcf	
enable-cnv-tracks	Permet d'activer ou de désactiver la génération de fichiers bigWig et GFF.	--enable-cnv-tracks	true/false
enable-combinegvcfs	Permet d'activer ou de désactiver la fusion de fichiers VCF.	--enable-combinegvcfs	true/false
enable-joint-genotyping	Pour activer la fusion de fichiers VCF, mettez la valeur à « true » (activée).	--enable-joint-genotyping	true/false
enable-multi-sample-gvcf	Permet d'activer ou de désactiver la génération d'un fichier VCF multiéchantillons. Si la valeur est « true » (activée), le fichier d'entrée doit être un fichier VCF combiné.	--enable-multi-sample-gvcf	true/false
enable-sv	Permet d'activer ou de désactiver la fonction de définition de variants structurels. La valeur par défaut est « false » (désactivée).	--enable-sv	true/false

Nom	Description	Équivalent dans la ligne de commande	Valeurs
enable-vlrd	Permet d'activer ou de désactiver la détection virtuelle de lectures longues (VLRD).	--enable-vlrd	true/false
panel-of-normals	Le chemin d'accès menant au fichier VCF du panel de référence.	--panel-of-normals	
pedigree-file	Option propre à la définition combinée. Le chemin d'accès vers un fichier PED de généalogie contenant une description structurée des relations familiales entre les échantillons. Le fichier de généalogie peut contenir des trios. Seuls les fichiers de généalogie contenant des trios sont pris en charge.	--pedigree-file	
qc-snp-DeNovo-quality-threshold	Seuil de dénombrement et de production de rapport pour les variants SNP (polymorphisme mononucléotidique) de novo.	--qc-snp-DeNovo-quality-threshold	
qc-indel-DeNovo-quality-threshold	Le seuil de dénombrement et de production de rapport pour les variants INDEL de novo.	--qc-indel-DeNovo-quality-threshold	
sv-call-regions-bed	Spécifie un fichier BED contenant un ensemble de régions à définir. Le fichier peut être compressé en format gzip ou bzip.	--sv-call-regions-bed	
sv-denovo-scoring	Permet d'activer ou de désactiver le score de qualité de novo pour la définition diploïde combinée de variants structurels. Un fichier de généalogie doit également être fourni.	--sv-denovo-scoring	
sv-exome	Lorsque la valeur est « true » (activée), la fonction de définition de variants est configurée pour les entrées de séquençage ciblées, y compris en désactivant les filtres de grandes profondeurs. La valeur par défaut est « false » (désactivée).	--sv-exome	true/false
sv-output-contigs	Mettez la valeur « true » (activée) pour produire des séquences de contigs assemblées dans un fichier VCF de sortie. La valeur par défaut est « false » (désactivée).	--sv-output-contigs	true/false
sv-region	Limite l'analyse à une région spécifiée du génome aux fins de débogage. L'option peut être spécifiée plusieurs fois pour créer une liste de régions.	--sv-region	La valeur doit être dans le format « chr:startPos-endPos ».
variant	Le chemin d'accès vers un fichier gVCF. L'option *--variant* peut être utilisée plusieurs fois sur la ligne de commande, une fois par fichier gVCF. Jusqu'à 500 fichiers gVCF peuvent être pris en charge.	--variant	
variant-list	Le chemin d'accès vers un fichier contenant une liste de fichiers gVCF d'entrée, un par ligne, à combiner.	variant-list	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
vc-decoy-contigs	Le chemin d'accès vers une liste de contigs, séparés par des virgules, à sauter pendant la définition de variants.	--vc-decoy-contigs	
vc-emit-ref-confidence	Pour activer la génération d'un fichier gVCF de paires de bases, mettez la valeur « BP_RESOLUTION ». Pour activer la génération d'un fichier gVCF à bandes, mettez la valeur « GVCF ».	--vc-emit-ref-confidence	BP_RESOLUTION/GVCF
vc-enable-baf	Permet d'activer ou de désactiver le fichier de sortie de fréquence d'allèle B. La valeur par défaut est « true » (activée).	--vc-enable-baf	
vc-enable-decoy-contigs	Permet d'activer ou de désactiver la définition de variants sur des contigs de leurres. La valeur par défaut est « false » (désactivée).	--vc-enable-decoy-contigs	true/false
vc-enable-gatk-acceleration	Permet d'activer ou de désactiver l'exécution de la fonction de définition de variants en mode GATK.	--vc-enable-gatk-acceleration	true/false
vc-enable-orientation-bias-filter	Permet d'activer ou de désactiver le filtre de biais d'orientation.	--vc-enable-orientation-bias-filter	true/false
vc-enable-phasing	Permet aux variants d'être mis en phase lorsque cela est possible. La valeur par défaut est « true » (activée).	--vc-enable-phasing	true/false
vc-enable-roh	Permet d'activer ou de désactiver la fonction de définition des régions d'homozygotie (ROH) et sa sortie. La valeur par défaut est « true » (activée).	--vc-enable-roh	
vc-forcegt-vcf	Force le génotypage pour la définition de petits variants germinaux. Un fichier .vcf ou .vcf.gz contenant une liste de petits variants est requise.	--vc-forcegt-vcf	Le fichier .vcf ou vcf.gz spécifiant les petits variants dont le génotypage est forcé.
vc-frd-beta-max	La fréquence allélique maximale pour l'hypothèse relative aux lectures étrangères.	--vc-frd-beta-max	
vc-gvcf-gq-bands	Définit les bandes de qualité du génotype (GQ) aux fins de production du fichier gVCF. Les valeurs par défaut sont « 10, 20, 30, 40, 60, 80 ».	--vc-gvcf-gq-bands	
vc-hard-filter	L'expression booléenne permettant de filtrer les définitions de variants. L'expression par défaut est : DRAGENHardQUAL:all: QUAL < 10.4139;LowDepth:all: DP < 1		Les paramètres de l'expression peuvent comprendre « QD », « MQ », « FS », « MQRankSum », « ReadPosRankSum », « QUAL », « DP » et « GQ ».
vc-limit-genomecov-output	Limite la taille du fichier .genomecov.bed généré. La valeur par défaut est « false » (désactivée).	--vc-limit-genomecov-output	true/false

Nom	Description	Équivalent dans la ligne de commande	Valeurs
vc-max-alternate-alleles	Le nombre maximal d'allèles ALT produits dans un fichier VCF ou gVCF. La valeur par défaut est « 6 ».	--vc-max-alternate-alleles	
vc-max-reads-per-active-region	Le nombre maximal de lectures par région active pour le sous-échantillonnage. La valeur par défaut est « 10 000 ».	--vc-max-reads-per-active-region	
vc-max-reads-per-active-region-mito	Le nombre maximal de lectures couvrant une région active donnée pour la définition de petits variants mitochondriaux. La valeur par défaut est « 40 000 ».	--vc-max-reads-per-active-region-mito	
vc-max-reads-per-raw-region	Le nombre maximal de lectures par région brute pour le sous-échantillonnage. La valeur par défaut est « 30 000 ».	--vc-max-reads-per-raw-region	
vc-max-reads-per-raw-region-mito	Le nombre maximal de lectures couvrant une région brute spécifiée pour la définition de petits variants mitochondriaux. La valeur par défaut est « 40 000 ».	--vc-max-reads-per-raw-region-mito	
vc-min-base-qual	La qualité des bases minimale à considérer pour la définition de variants. La valeur par défaut est « 10 ».	--vc-min-base-qual	
vc-min-call-qual	La qualité minimale de définition de variants pour produire une définition. La valeur par défaut est « 3 ».	--vc-min-call-qual	
vc-min-read-qual	La qualité de lecture (MAPQ) minimale à considérer pour la définition de variants. La valeur par défaut est « 10 ».	--vc-min-read-qual	
vc-min-reads-per-start-pos	Le nombre minimal de lectures par position de départ pour le sous-échantillonnage. La valeur par défaut est « 10 ».	--vc-min-reads-per-start-pos	
vc-min-tumor-read-qual	La qualité minimale de lecture (MAPQ) de tumeur à considérer pour la définition de variants.		
vc-orientation-bias-filter-artifacts	Le type d'artéfacts à filtrer. Un artéfact (ou un artéfact et son complément inverse) ne peut pas apparaître plus d'une fois dans la liste.	--vc-orientation-bias-filter-artifacts	C/T, G/T ou C/T, G/T, C/A
vc-remove-all-soft-clips	Si la valeur est « true » (activée), la fonction de définition de variants n'utilise pas de bases coupées présentes dans l'alignement des lectures pour déterminer les variants.	--vc-remove-all-soft-clips	true/false
vc-roh-blacklist-bed	Le fichier BED de régions à ignorer pour l'algorithme ROH (régions d'homozygotie).	--vc-roh-blacklist-bed	
vc-target-bed	Le fichier BED de régions cibles.	--vc-target-bed	
vc-target-bed-padding	Peut être utilisée pour enrichir toutes les régions cibles de fichier BED ayant la valeur spécifiée (facultatif). Si une valeur est spécifiée, elle est utilisée par la fonction de définition de petits variants.	--vc-target-bed-padding	

Nom	Description	Équivalent dans la ligne de commande	Valeurs
vc-target-coverage	La couverture cible pour le sous-échantillonnage. La valeur par défaut est « 500 » pour le mode germinal et « 50 » pour le mode somatique.	--vc-target-coverage	
vc-target-coverage-mito	Le nombre maximal de lectures dont la position de départ chevauche toute autre position donnée pour la définition de petits variants mitochondriaux. La valeur par défaut est « 40 000 ».	--vc-target-coverage	
vc-tlod-filter-threshold	Définit le seuil du filtre TLOD (limite de détection de tumeur). La valeur par défaut est « 17,5 » pour le mode relatif aux échantillons de tumeur et de référence et « 6,5 » pour le mode relatif aux échantillons de tumeur seulement.	--vc-tlod-filter-threshold	

Options de la fonction de détection des expansions de répétition

Les options suivantes peuvent être réglées dans la section RepeatGenotyping du fichier de configuration ou dans la ligne de commande. Pour plus de renseignements, consultez la section *Détection des expansions de répétition avec la méthode de détection ExpansionHunter* à la page 82.

Nom du paramètre	Description	Équivalent dans la ligne de commande	Valeurs
enable	Active ou désactive la détection des expansions de répétition.	--repeat-genotype-enable	true/false
specs	Le chemin d'accès complet vers le fichier JSON qui contient le catalogue des variants de répétitions (les spécifications) décrivant le locus à définir.	--repeat-genotype-specs	

Assistance technique

Pour obtenir de l'assistance technique, communiquez avec l'assistance technique d'Illumina.

Site Web : www.illumina.com
Courriel : techsupport@illumina.com

Numéros de téléphone de l'assistance clientèle d'Illumina

Région	Numéro sans frais	Numéro régional
Amérique du Nord	+(1) 800 809 4566	
Allemagne	+49 8001014940	+49 8938035677
Australie	+(1) 800 775 688	
Autriche	+43 800006249	+43 19286540
Belgique	+32 80077160	+32 34002973
Chine	400 066 5835	
Corée du Sud	+82 80 234 5300	
Danemark	+45 80820183	+45 89871156
Espagne	+34 911899417	+34 800300143
Finlande	+358 800918363	+358 974790110
France	+33 805102193	+33 170770446
Hong Kong, Chine	800960230	
Irlande	+353 1800936608	+353 016950506
Italie	+39 800985513	+39 236003759
Japon	0800 111 5011	
Norvège	+47 800 16836	+47 21939693
Nouvelle-Zélande	0800 451 650	
Pays-Bas	+31 8000222493	+31 207132960
Royaume-Uni	+44 8000126019	+44 2073057197
Singapour	+(1) 800 579 2745	
Suède	+46 850619671	+46 200883979
Suisse	+41 565800000	+41 800200442
Taïwan, Chine	00806651752	
Autres pays	+44 1799 534 000	

Fiches signalétiques (SDS) : disponibles sur le site Web d'Illumina à l'adresse support.illumina.com/sds.html.

Documentation sur les produits : disponible en téléchargement sur le site support.illumina.com.

Document n° 1000000111887 v00 FRA

Illumina
5200 Illumina Way
San Diego, Californie 92122 États-Unis
+ (1) 800 809 ILMN (4566)
+ (1) 858 202 4566 (en dehors de l'Amérique du Nord)
techsupport@illumina.com
www.illumina.com

Destiné à la recherche uniquement.
Ne pas utiliser dans le cadre d'examens diagnostiques.

illumina®