illumina®

DRAGEN for Illumina DNA Prep with Enrichment Dx on NextSeq 550Dx

Guide de l’utilisateur de l’application

PROPRIÉTÉ D’ILLUMINA
Document n° 200025238 v00
Février 2023
DESTINÉ AU DIAGNOSTIC IN VITRO UNIQUEMENT.

Historique des modifications

Document	Date	Description de la modification
200025238 v00	Février 2023	Publication initiale.

Table des matières

Présentation

L'application DRAGEN for Illumina DNA Prep with Enrichment Dx (DRAGEN pour IDPE Dx) est utilisée pour planifier et effectuer une analyse secondaire des bibliothèques IDPE Dx générées pour le séquençage sur le NextSeq 550Dx.

DRAGEN pour IDPE Dx prend en charge le séquençage jusqu'à l'analyse lorsqu'il est utilisé avec la préparation de bibliothèques Illumina DNA Prep with Enrichment Dx, NextSeq 550Dx et Illumina DRAGEN Server for NextSeq 550Dx.

Méthodes d'analyse

DRAGEN pour IDPE Dx effectue le démultiplexage, la génération FASTQ, la cartographie de lecture, l'alignement sur un génome de référence et la définition de variante en fonction des flux de travail sélectionnés :

- Génération des fichiers FASTQ
- Génération de fichiers FASTQ et VCF pour les variants germinaux
- Génération de fichiers FASTQ et VCF pour les variants somatiques

REMARQUE La compression ORA peut être utilisée avec les trois flux de travail. La compression DRAGEN ORA est un logiciel de compression des données sans aucune perte qui crée un fichier avec une extension Original Read Archive (*.ora). Le format ora est un format de compression basé sur la référence pour les fichiers FASTQ ; il est conçu pour une compression/décompression très rapide et un taux de compression élevé.

Génération des fichiers FASTQ

Les séquences assemblées sont écrites dans des fichiers FASTQ par échantillon. Les fichiers FASTQ sont des fichiers texte qui contiennent des données de séquençage et des scores de qualité pour un seul échantillon. Pour chaque échantillon, des fichiers FASTQ distincts sont générés par ligne de Flow Cell, par lecture de séquençage. Le nom de l'échantillon tel qu'il est spécifié lors de la configuration de la série est inclus dans le nom du fichier FASTQ. Les fichiers FASTQ constituent les principales données d'entrée pour l'alignement. La première étape de la génération de fichiers FASTQ est le démultiplexage. Le démultiplexage attribue des amplifiats qui franchissent le filtre vers un échantillon en comparant chaque séquence de lecture d'index aux séquences d'indexage définies pour la série. Aucune valeur de qualité n'est prise en compte lors de cette étape. Les lectures d'index sont identifiées en suivant les étapes ci-dessous :

- Les échantillons sont numérotés en commençant par 1, selon l'ordre dans lequel ils sont classés pour la série.
- Le numéro d'échantillon 0 est réservé aux amplifiats qui n'ont pas été assignés à un échantillon.
- Les amplifiats sont assignés à un échantillon lorsque la séquence d'indexage est identique ou lorsqu'il y a une seule non-correspondance par lecture d'index.

Le logiciel inclut la compression ORA pour compresser les fichiers FASTQ. Ce format peut être activé de manière facultative. Lors de l'utilisation du format ORA (*.ora), la somme de contrôle md5 du contenu FASTQ est conservée après un cycle de compression et de décompression pour assurer une compression sans perte.

Alignement et cartographie de l'ADN

Après la génération des fichiers FASTQ, les lectures sont cartographiées et alignées sur un génome de référence. 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 une non-correspondance 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. L'algorithme est accéléré sur les cartes réseau de portes programmables in situ (FPGA) de DRAGEN. Le génome de référence utilisé dans l'application est créé à partir du fichier UCSC hg19 FASTA avec l'option DRAGEN pour créer une table de hachage alt-aware basée sur un fichier liftover.

Définition de variants germinaux DRAGEN

Le DRAGEN Germline Small Variant Caller prend les lectures d'ADN cartographiées et alignées comme entrée et définit les polymorphismes nucléotidiques uniques (SNP) et les insertions ou délétions (indels) grâce à une combinaison de détection par colonne et d'assemblage local *de novo* d'haplotypes. Pour activer DRAGEN Germline Small Variant Caller, sélectionnez le flux de travail pour les variants germinaux.

La définition de variants germinaux est généralement utilisée pour les échantillons germinaux où la ploïdie est connue pour être double. Les régions de référence définissables sont d'abord identifiées avec une couverture d'alignement suffisante. Dans ces régions de référence, une analyse rapide des lectures triées identifie les régions actives, qui sont centrées sur des colonnes d'empilement avec des

preuves d'un variant. Les régions actives sont élargies avec suffisamment de contexte pour couvrir un contenu important et non référencé à proximité. S'il y a des preuves d'indels, les régions actives reçoivent un élargissement supplémentaire.

Les lectures alignées sont découpées dans chaque région active et assemblées dans un graphique De Bruijn. Les bords des lectures tronquées sont pondérés par le nombre d'observations, avec la séquence de référence comme pilier. Après un nettoyage et une simplification du graphique, tous les chemins source-puits sont extraits en tant qu'haplotypes candidats. Chaque haplotype est aligné selon Smith-Waterman sur le génome de référence pour identifier les variants qu'il représente. Cet ensemble d'événements peut être complété par une détection basée sur la position. Pour chaque paire lecture-haplotype, la probabilité P(r|H) d'observer la lecture, en supposant que l'haplotype est le véritable échantillon de départ, est estimée à l'aide d'un modèle de Markov caché (MMC).

En balayant par position de référence sur la région active, les génotypes candidats sont formés à partir de combinaisons diploïdes d'événements variants (SNP ou indels). Pour chaque événement (y compris la référence), la probabilité conditionnelle P(r|e) d'observer chaque lecture qui se chevauche est estimée comme le maximum P(r|H) pour les haplotypes supportant l'événement. Celles-ci sont combinées dans la probabilité conditionnelle P(r|e1e2) pour un génotype (paire d'événements) et multipliées pour donner la probabilité conditionnelle P(R|e1e2) d'observer l'empilement de lecture complet. En utilisant la formule de Bayes, la probabilité postérieure P(e1e2|R) de chaque génotype diploïde est calculée et le gagnant est défini.

DRAGEN pour IDPE Dx applique le filtrage automatique. Consultez la section *Annotations du fichier VCF pour le flux de travail germinal* à la page 14 pour plus d'informations.

Définition des variants somatiques DRAGEN

Le DRAGEN Somatic Small Variant Caller prend les lectures d'ADN cartographiées et alignées comme entrée et défini les SNV et les indels grâce à un assemblage local *de novo* d'haplotypes dans une région active. Pour activer DRAGEN Somatic Small Variant Caller, sélectionnez une application de variants somatiques.

La définition de variants somatiques est généralement utilisée pour les échantillons tumoraux. Avec ce flux de travail, DRAGEN ne fait aucune hypothèse de ploïdie, ce qui permet la détection d'allèles basse fréquence. Pour les loci avec une couverture allant jusqu'à 100x dans l'échantillon tumoral, DRAGEN a un seuil de détection à des fréquences d'allèles de variants de 5 %. La limite évolue avec une profondeur croissante sur une base par locus et de moitié chaque fois que la couverture double au-delà de 100x. Les régions de référence définissables sont d'abord identifiées avec une couverture d'alignement suffisante. Dans ces régions de référence, une analyse des lectures triées identifie les régions actives, qui sont centrées sur des colonnes d'empilement avec des preuves d'un variant dans les lectures de tumeur. Les régions actives sont élargies avec suffisamment de contexte pour couvrir un contenu important et non référencé à proximité. S'il y a des preuves d'indels, les régions actives reçoivent un élargissement supplémentaire.

Les lectures alignées sont découpées dans chaque région active et assemblées dans un graphique De Bruijn. Les bords des lectures tronquées sont pondérés par le nombre d'observations, avec la séquence de référence comme pilier. Après un nettoyage et une simplification du graphique, tous les chemins source-puits sont extraits en tant qu'haplotypes candidats. Chaque haplotype est aligné selon Smith-Waterman sur le génome de référence pour identifier les variants qu'il représente. Pour chaque paire lecture-haplotype, la probabilité P(r|H) d'observer la lecture, en supposant que l'haplotype est le véritable échantillon de départ, est estimée à l'aide d'un modèle de Markov caché (MMC).

Pour déterminer le score limite de détection tumorale (TLOD), DRAGEN Somatic Small Variant Caller analyse d'abord par position de référence pour chaque événement somatique candidat ainsi que l'événement de référence sur la région active. La probabilité conditionnelle P(r|e) d'observer chaque lecture qui se chevauche est estimée comme le maximum P(r|H) pour les haplotypes supportant l'événement. Celles-ci sont combinées dans la probabilité conditionnelle P(r|E) pour une hypothèse d'événement, E, impliquant un mélange de l'allèle somatique de référence et candidat sur une gamme de fréquences d'allèles possibles et multipliées pour donner la probabilité conditionnelle P(R|E) d'observer tout l'empilement de lecture. À partir de là, un score TLOD est calculé comme la preuve qu'un allèle ALT est présent dans l'échantillon de tumeur à un locus donné.

DRAGEN pour IDPE Dx applique le filtrage automatique. Consultez la section *Annotations du fichier VCF pour le flux de travail somatique* à la page 17 pour plus d'informations.

Créer une série planifiée

Utilisez les étapes suivantes pour configurer une série dans Illumina Run Manager soit sur le NextSeq 550Dx ou à l'aide d'un navigateur sur un ordinateur en réseau. Utilisez un navigateur sur un ordinateur en réseau si vous souhaitez importer des données de l'échantillon. Consultez le Guide du logiciel Illumina Run Manager pour NextSeq 550Dx (document n° 200025239) pour obtenir des instructions sur l'accès à Illumina Run Manager partir d'un ordinateur en réseau.

Il existe deux façons différentes de créer une nouvelle série planifiée :

- **Import Run** (Importer une série)—Utilisez une feuille d'échantillon d'une série existante comme modèle pour une nouvelle série. Consultez le Guide du logiciel Illumina Run Manager pour NextSeq 550Dx (document n° 200025239) pour plus d'informations sur la façon d'importer une série.
- **Create Run** (Créer une série)—Saisissez manuellement les paramètres de la série. Les instructions suivantes expliquent comment créer une série.

REMARQUE Les champs de saisie obligatoires dans l'interface utilisateur sont indiqués par un astérisque (*).

Application

1. Dans l'onglet Planned (Planifiée) de l'écran Runs (Séries), sélectionnez **Create Run** (Créer une série).
2. Sélectionnez l'application DRAGEN pour Illumina DNA Prep with Enrichment Dx, puis sélectionnez **Next** (Suivant).

Paramètres de la série

1. Sur l'écran Run Settings (Paramètres de la série), entrez un nom unique pour la série. Le nom de la série identifie la série depuis le séquençage jusqu'à l'analyse.
2. **[Facultatif]** Entrez une description de série pour identifier davantage la série.
3. Sélectionnez la ou les trousse(s) d'adaptateurs d'index utilisées pendant la préparation de la bibliothèque.
4. Vérifiez la longueur de lecture et modifiez-la, si nécessaire. La lecture 1 et la lecture 2 ont une valeur par défaut de 151 cycles. L'index 1 et l'index 2 ont une valeur fixe de 10 cycles et ne peuvent pas être modifiés.
5. **[Facultatif]** Entrez un ID de tube de bibliothèque.
6. Sélectionnez **Next** (Suivant).

Données de l'échantillon

Les données d'échantillon comprennent l'ID d'échantillon, la position du puits (position du puits de la plaque d'index) et le nom de la bibliothèque. Lors de l'utilisation de l'index A&B, la position du puits comprend également l'identifiant de la plaque.

Il existe deux façons de saisir des données d'échantillon :

- **Import Samples** (Importer des échantillons)—Utilisez un fichier modèle disponible pour le téléchargement sur l'écran Sample Data (Données d'échantillon).
- **Manually** (Manuellement)—Saisissez les données d'échantillon directement dans le tableau sur l'écran Sample Data (Données d'échantillon).

Importer des échantillons

Lors de la planification d'une série de séquençage à l'aide d'un navigateur sur un ordinateur en réseau, un fichier modèle (`*.csv`) est disponible en téléchargement sur l'écran Sample Data (Données de l'échantillon). Le fichier modèle n'est pas disponible en téléchargement lors de l'accès à Illumina Run Manager via le logiciel du système d'exploitation de NextSeq 550Dx. Pour saisir des données d'échantillon à l'aide de la fonctionnalité Import Samples (Importer des échantillons), procédez comme suit.

REMARQUE Effectuez les étapes pour les paramètres de la série avant de continuer.

1. Sélectionnez **Download Template** (Télécharger le modèle) pour télécharger un fichier CSV vierge.
2. À partir du fichier modèle, saisissez les données d'échantillon, puis enregistrez le fichier. Le nom de la bibliothèque est facultatif.

REMARQUE Lors de l'utilisation de l'index A&B, les données pour la colonne B doivent inclure la position de la plaque et du puits (position du puits de la plaque d'index). Exemple : A-A01, A-A02, A-A03.

3. Sélectionnez **Import Samples** (Importer des échantillons) et naviguez jusqu'au fichier modèle contenant les informations sur les données d'échantillon de l'étape précédente.
4. Sélectionnez **Open** (Ouvrir), **Proceed** (Continuer), puis **Next** (Suivant).

REMARQUE Une modification de l'ID d'échantillon avant de sélectionner Next (Suivant) peut entraîner une erreur. Terminez la configuration de la série avant d'apporter des modifications afin d'éviter les erreurs.

Saisir les échantillons manuellement

Utilisez le tableau de l'écran Sample Data (Données de l'échantillon) pour saisir manuellement les données de l'échantillon.

1. Saisissez un ID d'échantillon unique dans le champ Sample ID (ID d'échantillon).
2. Utilisez **Well Position** (Position du puits) (Index A ou Index B) ou **Plate - Well Position** (Plaque - Position du puits) (Index A&B) pour sélectionner l'index associé aux échantillons.
   Les champs Index i7, Index 1, Index i5 et Index 2 se remplissent automatiquement.
3. **[Facultatif]** Entrez le nom de la bibliothèque.
4. Ajoutez des lignes et répétez les étapes 1 à 3 si nécessaire jusqu'à ce que tous les échantillons aient été ajoutés au tableau. Vous pouvez ajouter plusieurs lignes simultanément en saisissant d'abord le nombre de lignes à ajouter, puis en sélectionnant l'icône +. Vous pouvez également supprimer des lignes en sélectionnant la case à côté du numéro de la ligne, puis en cliquant sur l'icône de la corbeille.
5. Sélectionnez **Next** (Suivant).

Paramètres de l'analyse

1. Sélectionnez le flux de travail d'analyse souhaité :
   • Génération de fichiers FASTQ
   • Génération de fichiers FASTQ et VCF pour un flux de travail germinal (fichier de manifeste requis)
   • Génération de fichiers FASTQ et VCF pour un flux de travail somatique (fichier de manifeste requis)
2. **[Facultatif]** L'option **Generate ORA compressed FASTQs** (Générer des FASTQ compressés ORA) est activée par défaut. La compression ORA des FASTQ compresse sans perte les fichiers FASTQ jusqu'à 5 fois par rapport à la compression fastq.gz. Décochez l'option **Generate ORA compressed FASTQs** (Générer des FASTQ compressés ORA) si vous préférez obtenir des données non compressées (fastq.gz).
3. Pour les flux de travail germinaux et somatiques, un fichier manifeste est requis. Utilisez le menu déroulant **Manifest File Selection** (Sélection du fichier de manifeste) pour sélectionner un fichier de manifeste. Le manifeste est un fichier BED délimité par des tabulations (`*.bed`) qui spécifie les noms et les emplacements des régions de référence ciblées. Pour plus d'informations, consultez la section *Fichier de manifeste* à la page 9.
4. **[Facultatif]** Pour les flux de travail somatiques, utilisez le menu déroulant **Noise File Selection** (Sélection du fichier de bruit) pour sélectionner un fichier de bruit systématique.
   Un fichier BED (`*.bed.gz`) avec un niveau de bruit spécifique au site peut être spécifié pour filtrer le bruit systématique. Pour plus d'informations, consultez la section *Filtrage du bruit (Facultatif)* à la page 9.
5. Sélectionnez **Next** (Suivant).

Série Examiner

1. Sur l'écran Review (Examiner), passez en revue les informations saisies pour Run Settings (Paramètres de la série), Sample Data (Données de l'échantillon) et Analysis Settings (Paramètre de l'analyse).
2. Sélectionnez **Save** (Enregistrer).
   La série est enregistrée dans l'onglet Planned (Planifiée) de l'écran Runs (Séries).

Paramètres

Pour afficher ou modifier les paramètres de l'application DRAGEN pour IDPE Dx, sélectionnez d'abord l'icône Applications sur l'écran principal. Sélectionnez ensuite l'application que vous souhaitez afficher ou modifier. Un compte administrateur est requis pour pouvoir modifier les paramètres.

Configuration

L'écran de configuration affiche les paramètres d'application suivants :

- **Library Prep Kits** (Trousses de préparation de bibliothèque) : affiche la trousse de préparation de bibliothèque par défaut pour l'application. Ce paramètre ne peut pas être modifié.
- **Index Adapter Kits** (Trousses d'adaptateur d'index) : affiche la trousse d'adaptateur d'index par défaut pour l'application. Ce paramètre ne peut pas être modifié.
- **Read lengths** (Longueurs de lecture) : les longueurs de lecture sont définies sur 151 pour l'application par défaut, mais peuvent être modifiées lors de la création de la série.
- **Manifest and Noise Files** (Fichiers de manifeste et de bruit) : téléchargez et modifiez les paramètres des fichiers de manifeste et de bruit.
  - Sélectionnez **Upload File** (Télécharger le fichier) pour télécharger les fichiers à utiliser dans l'analyse.
  - Sélectionnez le bouton radio **Default** (Par défaut) pour définir le fichier comme fichier de manifeste ou de bruit par défaut sélectionné lors de la création de la série lorsque l'application est sélectionnée.
  - Cochez la case **Enabled** (Activé) pour définir le fichier à afficher dans le menu déroulant lors de la création de la série.

Autorisations

Utilisez les cases à cocher sur l'écran Permissions (Autorisations) pour gérer l'accès des utilisateurs à l'application.

Fichier de manifeste

Lors de l'utilisation de DRAGEN pour IDPE Dx, un fichier de manifeste est requis pour les flux de travail suivants :

- Génération de fichiers FASTQ et VCF pour un flux de travail germinal
- Génération de fichiers FASTQ et VCF pour un flux de travail somatique

Le fichier de manifeste est un fichier texte délimité par des tabulations utilisant le format BED (*.bed) qui spécifie les noms et les emplacements des régions de référence ciblées. La section principale du fichier de manifeste est la section Regions (Régions) et doit contenir les colonnes de données suivantes :

Colonne	Description
Nom	Nom unique spécifié par l'utilisateur pour la cible
Chromosome	Emplacement du chromosome (p. ex. chr10, chr5, etc.)
Départ	Index basé sur 1 pour la position de départ de la cible
Arrêt	Index basé sur 1 pour la position d'arrêt de la cible
Longueur de la sonde en amont	La longueur de la sonde en amont. Pour l'application DRAGEN pour IDPE Dx, elle doit être définie sur 0.
Longueur de la sonde en aval	La longueur de la sonde en aval. Pour l'application DRAGEN pour IDPE Dx, elle doit être définie sur 0.

REMARQUE Un format de fichier de manifeste valide est requis pour l'analyse. DRAGEN arrêtera l'analyse si le fichier de manifeste n'est pas valide.

Filtrage du bruit (Facultatif)

Le filtre de bruit systématique est disponible pour la définition de variants somatiques et peut être utilisé pour réduire les faux positifs en tenant compte du bruit spécifique au site. Le fichier de bruit systématique est généré en recueillant d'abord environ 50 échantillons normaux (de préférence spécifiques au panel, à la préparation de la bibliothèque et au séquenceur), puis la somme des fréquences d'allèles inférieures à 30 % sur chaque site avec une couverture suffisante est divisée par le nombre total d'échantillons (les fréquences d'allèles supérieures à 30 % sont supposées être des variants germinaux et non du bruit). Une fois les valeurs de bruit générées, les variants somatiques détectés sur ce site seront filtrés.

Le filtre peut être utilisé en mode Tumeur normale, mais est particulièrement utile pour les séries Tumor-Only (Tumeur uniquement) lorsqu'une tumeur normale correspondante n'est pas disponible. Le fichier de bruit systématique doit utiliser un fichier BED avec une extension de fichier (`*.bed.gz`) et doit inclure quatre colonnes : Chromosome (Chromosome), Start (Début), End (Fin) et les niveaux de bruit spécifiques au site pour chaque ligne. Le filtrage du bruit systématique est facultatif.

Sorties d'analyse

Les séries en cours sont affichées dans l'onglet Active (Actif). Les séries terminées sont affichées dans l'onglet Completed (Terminé). DRAGEN pour IDPE Dx crée un dossier d'analyse nommé de manière unique pour chaque analyse, qui est distinct du dossier contenant les données de séquençage. Le dossier d'analyse comprend les informations suivantes :

- Fichier de manifeste utilisé
- Version du logiciel
- ID des échantillons
- Nombre total de lectures alignées
- Pourcentage de lectures alignées par échantillon
- Nombre de SNV définis par échantillon
- Nombre d'indels définis par échantillon
- Statistiques de couverture

Fichiers de sortie d'analyse

L'emplacement du dossier d'analyse est spécifié par le paramètre External Storage for Analysis Results (Stockage externe pour les résultats d'analyse). Consultez le Guide du logiciel Illumina Run Manager pour NextSeq 550Dx (document n° 200025239) pour plus d'informations sur le paramètre External Storage for Analysis Results (Stockage externe pour les résultats d'analyse).

Sur l'écran Run Details (Détails de la série), le champ External Location (Emplacement externe) fournit le chemin d'accès pour le séquençage des données. Le nom unique du dossier d'analyse est fourni dans le champ Analysis Output Folder (Dossier de sortie de l'analyse) sur l'écran Run Details (Détails de la série). Les fichiers exacts générés dépendent du flux de travail d'analyse utilisé. Les fichiers de sortie de l'analyse suivants sont générés par l'application.

REMARQUE Si une erreur de limitation de la longueur maximale du chemin d'accès au fichier se produit lors de l'accès aux fichiers de sortie de l'analyse, essayez de déplacer le fichier vers un emplacement de chemin d'accès plus court ou utilisez une méthode différente pour ouvrir le fichier.

Fichier de sortie	Description
Rapport récapitulatif des variants (`*.pdf`)	Contient un résumé des informations sur le fichier, les versions logicielles, les informations sur les échantillons, les statistiques du niveau de lecture et les résumés des SNV, insertions, délétions et de la couverture. Seuls les flux de travail des variants germinaux et somatiques produisent un rapport sur les variants.
FASTQ (`*.fastq.gz` ou `*.fastq.ora`)	Fichiers intermédiaires contenant les définitions de bases dont la qualité est notée. Les fichiers FASTQ constituent les principales données d'entrée pour l'étape de l'alignement. Lorsque ORA Compression est sélectionné, l'extension de fichier `*.fastq.ora` est utilisée.
Fichiers d'alignement BAM (`*.bam`)	Contiennent les lectures alignées pour un échantillon donné.
Fichiers VCF du génome (`*.gvcf.gz`)	Contiennent le génotype pour chacune des positions, qu'elles soient définies à titre de variant ou de référence.
Fichiers VCF (`*.vcf.gz`)	Contiennent les variants appelés à chacune des positions.
Rapport d'indicateurs de la série (`*.csv`)	Contient des indicateurs de qualité sur la série, y compris le rendement total non indexé et le score Q30.

Fichiers FASTQ

FASTQ (*.fastq.gz, *.fastq.ora) est un format de fichier texte qui contient les définitions de bases et les valeurs de qualité, par lecture. Chaque fichier contient les informations suivantes :

- L'identifiant de l'échantillon
- La séquence
- Les scores de qualité Phred sont identifiés dans un format codé ASCII + 33

L'identifiant de l'échantillon est formaté comme suit :

```
@Instrument:IDSérie:IDFlowCell:Ligne:Plaque:X:Y
NumLecture:MarqueurFiltre:0:NuméroÉchantillon
Exemple :
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAA9#:<#<;<<<????#=
```

Fichiers BAM

Un fichier BAM (*.bam) est la version binaire compressée d'un fichier SAM (sequence alignment map) utilisée pour représenter des séquences alignées, jusqu'à 128 Mo. Les fichiers BAM utilisent le format de dénomination de fichier `SampleName_S#.bam`. # est le numéro d'échantillon déterminé par l'ordre dans lequel les échantillons sont répertoriés pour la série. En mode multinœud, le S# est défini sur S1, quel que soit l'ordre de l'échantillon.

Les fichiers BAM contiennent une section d'en-tête et une section pour les alignements :

- **Header** (En-tête) : contient les renseignements sur la totalité du fichier, tels que le nom de l'échantillon, sa longueur et la méthode d'alignement. Les alignements dans la section des alignements sont associés aux renseignements spécifiques figurant dans la section d'en-tête.
- **Alignments** (Alignements) : contient le nom de la lecture, la séquence de lecture, la qualité de la lecture, les renseignements sur l'alignement et des balises personnalisées. Le nom de la lecture comprend le chromosome, les coordonnées de départ, la qualité de l'alignement et la chaîne du descripteur de correspondance.

La section des alignements comprend les renseignements suivants sur chacune des lectures ou des paires de lectures :

- AS : qualité de l'alignement des paires de bases appariées.
- RG : groupe de lecture, qui indique le nombre de lectures pour un échantillon spécifique.
- BC : marqueur du code à barres qui indique l'identifiant de l'échantillon démultiplexé associé à la lecture.
- SM : qualité de l'alignement à paire de bases unique.
- XC : chaîne du descripteur de correspondance.
- XN : marqueur du nom de l'amplicon, qui enregistre l'identifiant de l'amplicon associé à la lecture

Les fichiers d'index BAM (*.bam.bai) fournissent un index du fichier BAM correspondant.

Fichiers VCF

Les fichiers au format VCF (variant call format, *.vcf) contient des renseignements sur les variants que l'on trouve à des positions spécifiques dans un génome de référence.

L'en-tête du fichier VCF comprend la version du format de fichier VCF, la version du paramètre de définitions des variants et les annotations utilisées dans le reste du fichier. L'en-tête du fichier comprend aussi le fichier du génome de référence et le fichier BAM. La dernière ligne de l'en-tête est constituée d'en-têtes de colonnes pour les lignes de données. Chacune des lignes de données d'un fichier VCF contient des renseignements sur un seul variant.

Tableau 1 En-têtes du fichier VCF

En-tête	Description
CHROM	Chromosome du génome de référence. Les chromosomes apparaissent dans le même ordre que dans le fichier de référence FASTA.
POS	Position de la base unique du variant dans le chromosome de référence. Pour les variants mononucléotidiques (SNV), cette position est la base de référence avec le variant. Pour les indels, cette position est la base de référence immédiatement avant le variant.
ID	Le numéro rs (SNP de référence) pour le SNP obtenu à partir du fichier `dbSNP.txt` le cas échéant. S'il existe plusieurs numéros rs à cet emplacement, la liste est délimitée par des points-virgules. Si aucune entrée dbSNP n'existe à cette position, un marqueur de valeur manquante ('.') est utilisé.
REF	Le génotype de référence. Par exemple, une délétion d'un T unique est représentée TT comme référence et T comme alternative. Un variant à simple nucléotide de A à T est représenté A comme référence et T comme alternative.
ALT	Allèles qui diffèrent de la lecture de référence. Par exemple, une insertion d'un T unique est représentée A comme référence et AT comme alternative. Un variant à simple nucléotide de A à T est représenté A comme référence et T comme alternative.
QUAL	Score de qualité de l'échelle Phred attribué par le paramètre de définition des variants. Des scores supérieurs indiquent une confiance accrue dans le variant et une moindre probabilité d'erreur. Pour un score de qualité de Q, la probabilité estimée d'une erreur est de 10-(Q/10). Par exemple, l'ensemble d'appels Q30 est associé à un taux d'erreur de 0,1 %. De nombreux paramètres d'appel des variants attribuent des scores de qualité en fonction de leurs modèles statistiques, qui sont élevés par rapport au taux d'erreur observé.

Tableau 2 Annotations du fichier VCF pour le flux de travail germinal

En-tête	Description
FILTER (Filtre)	Si tous les filtres sont passés, PASS (Réussite) s'inscrit dans la colonne Filter (Filtre). Les entrées FILTER (Filtre) possibles incluent : • **DRAGENSnpHardQUAL** : appliqué si le score QUAL du variant SNP n'atteint pas le seuil • **DRAGENIndelHardQUAL** : appliqué si le score QUAL du variant indel n'atteint pas le seuil • **LowDepth** : site filtré, car la profondeur de couverture n'atteint pas le seuil • **LowGQ** : site filtré, car la qualité du génotype n'atteint pas le seuil • **PloidyConflict** : la définition de génotype du paramètre de définition de variant n'est pas compatible avec la ploïdie chromosomique • **base_quality** : site filtré, car la qualité de base médiane des lectures alt à ce locus n'atteint pas le seuil • **filtered_reads** : site filtré, car une trop grande fraction de lectures a été filtrée • **fragment_length** : site filtré, car la différence absolue entre la longueur médiane des fragments des lectures alt et la longueur médiane des fragments des lectures ref à ce locus dépasse le seuil • **low_depth** : site filtré, car la profondeur de lecture est trop faible • **low_frac_info_reads** : site filtré, car la fraction de lectures informatives est inférieure au seuil • **low_normal_depth** : site filtré, car la profondeur de lecture normale de l'échantillon est trop faible • **long_indel** : site filtré, car la longueur de l'indel est trop longue • **mapping_quality** : site filtré, car la qualité de cartographie médiane des lectures alt à ce locus n'atteint pas le seuil • **multiallelic** : site filtré, car plus de deux allèles alt passent la LOD de la tumeur • **non_homref_normal** : site filtré, car le génotype normal de l'échantillon n'est pas une référence homozygote • **no_reliable_supporting_read** : site filtré, car il n'existe aucune lecture somatique fiable • **panel_of_normals** : vu dans au moins un échantillon dans le panel des échantillons normaux vcf • **read_position** : site filtré, car la médiane des distances entre le début/la fin de la lecture et ce lieu est inférieure au seuil • **RMxNRepeatRegion** : site filtré, car tout ou partie de l'allèle variant est une répétition de la référence • **strand_artifact** : site filtré en raison d'une distorsion sévère du brin • **str_contraction** : site filtré en raison d'une erreur de PCR suspectée où l'allèle alt est une unité répétée de moins que la référence • **too_few_supporting_reads** : site filtré, car il y a trop peu de lectures de soutien dans l'échantillon de tumeur • **weak_evidence** : le score de variant somatique n'atteint pas le seuil

En-tête	Description
INFO	Les entrées INFO possibles incluent : • **AC** : comptage des allèles dans les génotypes pour chaque allèle ALT dans le même ordre que celui indiqué. • **AF** : fréquence des allèles pour chaque allèle ALT dans le même ordre que celui indiqué. • **AN** : nombre total d'allèles dans les génotypes appelés. • **DB** : adhésion dbSNP. • **FS** : valeur p à l'échelle de Phred utilisant le test exact de Fisher pour détecter la distorsion du brin. • **QD** : confiance/qualité par profondeur pour le variant. • **R2_5P_bias** : score basé sur la distorsion intersexuelle et la distance à partir de 5. • **SOR** : rapport de cotes symétrique du tableau de contingence 2x2 pour détecter la distorsion du brin. • **DP** : profondeur de lecture approximative (informative et non informative) ; certaines lectures peuvent avoir été filtrées en fonction de mapq, etc. • **END** : position d'arrêt de l'intervalle. • **FractionInformativeReads** : fraction de lectures informatives sur le nombre total de lectures. • **MQ** : qualité de mappage RMS. • **MQRankSum** : score Z du test de la somme des rangs de Wilcoxon des qualités de mappage de lecture Alt vs. Ref. • **ReadPosRankSum** : Z-score du test de somme des rangs de Wilcoxon du biais de position de lecture Alt vs. Ref. • **SOMATIC** : au moins un variant à cette position est un variant somatique.

En-tête	Description
FORMAT	La colonne Format dresse la liste des champs séparés par deux-points. Par exemple, GT:GQ. Les champs disponibles incluent : • **AD** : profondeurs alléliques (en ne comptant que les lectures informatives sur le total des lectures) pour les allèles ref et alt dans l'ordre indiqué. • **AF** : fractions d'allèles pour les allèles alt dans l'ordre indiqué. • **DP** : profondeur de lecture approximative (les lectures avec MQ=255 ou de mauvais appariements sont filtrées). • **F1R2** : nombre de lectures dans l'orientation de la paire F1R2 prenant en charge chaque allèle. • **F2R1** : nombre de lectures dans l'orientation de la paire F2R1 prenant en charge chaque allèle. • **GT** : génotype. 0 correspond à la base de référence, 1 correspond à la première entrée dans la colonne ALT, etc. La barre oblique (/) indique qu'aucun renseignement relatif à la mise en phase n'est disponible. • **MB** : statistiques de composants par échantillon pour détecter le biais intersexuel. • **PS** : informations d'identification de phase physique, où chaque ID unique au sein d'un échantillon donné (mais pas entre les échantillons) relie les enregistrements au sein d'un groupe de phase. • **SB** : statistiques de composants par échantillon qui comprennent le test exact de Fisher pour détecter la distorsion du brin. • **SQ** : qualité pour le variant somatique.
SAMPLE (Échantillon)	La colonne relative aux échantillons indique les valeurs précisées dans la colonne FORMAT.

Tableau 3 Annotations du fichier VCF pour le flux de travail somatique

En-tête	Description
FILTER (Filtre)	Si tous les filtres sont passés, PASS (Réussite) s'inscrit dans la colonne Filter (Filtre). Les entrées FILTER (Filtre) possibles incluent : • **base_quality** : site filtré, car la qualité de base médiane des lectures alt à ce locus n'atteint pas le seuil • **filtered_reads** : site filtré, car une trop grande fraction de lectures a été filtrée • **fragment_length** : site filtré, car la différence absolue entre la longueur médiane des fragments des lectures alt et la longueur médiane des fragments des lectures ref à ce locus dépasse le seuil • **low_depth** : site filtré, car la profondeur de lecture est trop faible • **low_frac_info_reads** : site filtré, car la fraction de lectures informatives est inférieure au seuil • **low_normal_depth** : site filtré, car la profondeur de lecture normale de l'échantillon est trop faible • **long_indel** : site filtré, car la longueur de l'indel est trop longue • **mapping_quality** : site filtré, car la qualité de cartographie médiane des lectures alt à ce locus n'atteint pas le seuil • **multiallelic** : site filtré, car plus de deux allèles alt passent la LOD de la tumeur • **non_homref_normal** : site filtré, car le génotype normal de l'échantillon n'est pas une référence homozygote • **no_reliable_supporting_read** : site filtré, car il n'existe aucune lecture somatique fiable • **panel_of_normals** : vu dans au moins un échantillon dans le panel des échantillons normaux vcf • **read_position** : site filtré, car la médiane des distances entre le début/la fin de la lecture et ce lieu est inférieure au seuil • **RMxNRepeatRegion** : site filtré, car tout ou partie de l'allèle variant est une répétition de la référence • **strand_artifact** : site filtré en raison d'une distorsion sévère du brin • **str_contraction** : site filtré en raison d'une erreur de PCR suspectée où l'allèle alt est une unité répétée de moins que la référence • **too_few_supporting_reads** : site filtré, car il y a trop peu de lectures de soutien dans l'échantillon de tumeur • **weak_evidence** : le score de variant somatique n'atteint pas le seuil • **systematic_noise** : site filtré sur la base de preuves de bruit systématique dans les échantillons normaux

En-tête	Description
INFO	Les entrées INFO possibles incluent : • **DP** : profondeur de lecture approximative (informative et non informative) ; certaines lectures peuvent avoir été filtrées en fonction de mapq, etc. • **END** : position d'arrêt de l'intervalle. • **FractionInformativeReads** : fraction de lectures informatives sur le nombre total de lectures. • **MQ** : qualité de mappage RMS. • **MQRankSum** : score Z du test de la somme des rangs de Wilcoxon des qualités de mappage de lecture Alt vs. Ref. • **ReadPosRankSum** : Z-score du test de somme des rangs de Wilcoxon du biais de position de lecture Alt vs. Ref. • **AQ** : score de bruit systématique. • **hotspot** : site somatique connu, utilisé pour augmenter la confiance dans la définition. • **SOMATIC** : au moins un variant à cette position est un variant somatique.

En-tête	Description
FORMAT	La colonne Format dresse la liste des champs séparés par deux-points. Par exemple, GT:GQ. Les champs disponibles incluent : • **AD** : profondeurs alléliques (en ne comptant que les lectures informatives sur le total des lectures) pour les allèles ref et alt dans l'ordre indiqué. • **AF** : fractions d'allèles pour les allèles alt dans l'ordre indiqué. • **DP** : profondeur de lecture approximative (les lectures avec MQ=255 ou de mauvais appariements sont filtrées). • **F1R2** : nombre de lectures dans l'orientation de la paire F1R2 prenant en charge chaque allèle. • **F2R1** : nombre de lectures dans l'orientation de la paire F2R1 prenant en charge chaque allèle. • **GP** : probabilités postérieures à l'échelle de Phred pour les génotypes tels que définis dans la spécification VCF. • **GQ** : qualité du génotype. • **GT** : génotype. 0 correspond à la base de référence, 1 correspond à la première entrée dans la colonne ALT, etc. La barre oblique (/) indique qu'aucun renseignement relatif à la mise en phase n'est disponible. • **MB** : statistiques de composants par échantillon pour détecter le biais intersexuel. • **PL** : probabilités normalisées à l'échelle Phred pour les génotypes tels que définis dans la spécification VCF. • **PRI** : probabilités antérieures à l'échelle de Phred pour les génotypes. • **PS** : informations d'identification de phase physique, où chaque ID unique au sein d'un échantillon donné (mais pas entre les échantillons) relie les enregistrements au sein d'un groupe de phase. • **SB** : statistiques de composants par échantillon qui comprennent le test exact de Fisher pour détecter la distorsion du brin. • **SQ** : qualité pour le variant somatique.
SAMPLE (Échantillon)	La colonne relative aux échantillons indique les valeurs précisées dans la colonne FORMAT.

Fichiers VCF du génome

Les fichiers VCF du génome (*.gvcf.gz) sont des fichiers qui respectent un ensemble de conventions pour la représentation de tous les sites du génome dans un format raisonnablement compact. Les fichiers gVCF comprennent tous les sites de la région d'intérêt dans un fichier unique, pour chacun des échantillons. Le fichier gVCF affiche « aucune définition » aux positions qui ne passent pas tous les filtres. L'indicateur de génotype (GT) ./. indique qu'il n'y a aucune définition.

Remise en file d'attente d'une analyse

Vous pouvez remettre une analyse en file d'attente si l'analyse a été arrêtée, si l'analyse a échoué ou si vous souhaitez réanalyser une série avec des paramètres différents. Pour remettre l'analyse dans la file d'attente, procédez comme suit :

1. Sur l'écran Run (Série), sélectionnez l'onglet Completed (Terminé), puis sélectionnez le nom de la série à réanalyser.
   Si la remise en file d'attente de l'analyse a déjà été effectuée, sélectionnez le nom de la série parente.
2. Sur l'écran Run Details (Détails de la série), après Sequencing Information (Informations sur le séquençage), sélectionnez **Requeue Analysis** (Remettre l'analyse en file d'attente).
3. Sélectionnez une option :
   - Requeue analysis with no changes (Remettre l'analyse en file d'attente sans modification)
   - Edit run settings and requeue analysis (Modifier les paramètres de la série et remettre l'analyse en file d'attente)
   - Requeue analysis with a different application (Remettre l'analyse en file d'attente avec une application différente)
4. Confirmez que l'emplacement où résident actuellement les données de séquençage est indiqué dans le champ **Sequencing data file path** (Chemin de fichier de données de séquençage).

REMARQUE Le chemin d'accès aux données de séquençage doit correspondre au chemin indiqué dans le paramètre External Storage for Analysis Results (Stockage externe pour les résultats d'analyse). Consultez le Guide du logiciel Illumina Run Manager pour NextSeq 550Dx (document n° 200025239) pour plus d'informations sur la modification du chemin de stockage externe.

5. Reanalysis Reason (Motif pour la réanalyse)
6. Sélectionnez **Requeue Analysis** (Remettre l'analyse en file d'attente).
7. Apportez les modifications souhaitées pour Run Settings (Paramètres de la série), Sample Data (Données d'échantillon) et Analysis Settings (Paramètres d'analyse).
8. Sélectionnez **Save** (Enregistrer). L'analyse commence à utiliser les paramètres d'analyse actuels.

Assistance technique

Pour une assistance technique, contactez le support techniqueIllumina.

Site Web :	www.illumina.com
E-mail :	techsupport@illumina.com

Fiches de données de sécurité (SDS) : disponibles sur le site Web d'Illumina à l'adresse support.illumina.com/sds.html.

Documentation sur les produits : disponible en téléchargement sur support.illumina.com.

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

IVD

EC REP

Illumina Netherlands B.V.
Steenoven 19
5626 DK Eindhoven
The Netherlands

Commanditaire australien
Illumina Australia Pty Ltd
Nursing Association Building
Level 3, 535 Elizabeth Street
Melbourne, VIC 3000
Australie

DESTINÉ AU DIAGNOSTIC IN VITRO UNIQUEMENT.

illumina®