

# DRAGEN for Illumina DNA Prep with Enrichment Dx

Documentation sur le produit NovaSeq 6000Dx

PROPRIÉTÉ D'ILLUMINA

Document n°200014776 v02

Septembre 2022

DESTINÉ AU DIAGNOSTIC IN VITRO UNIQUEMENT

## Historique des modifications

Document	Date	Description de la modification
200014776 v02	Septembre 2022	<p>Correction du format pour le fichier de manifeste du texte (*.txt) au format BED (*.bed) dans les instructions de création d'une analyse.</p> <p>Correction des fichiers VCF de consensus en fichiers VCF dans la section de sortie d'analyse.</p>
200014776 v01	Août 2022	<p>Ajout :</p> <p>Section Paramètres</p> <p>Section Filtrage de bruit systémique</p> <p>Mise à jour des instructions de création d'exécution pour inclure plus de détails.</p> <p>Correction des fautes de frappe et de grammaire.</p> <p>Précision que les instructions sont destinées à l'application lorsqu'elle est utilisée avec l'instrument NovaSeq 6000Dx.</p> <p>Mise à jour des informations sur le contenu du fichier de sortie VCF.</p>
200014776 v00	Mars 2022	Publication initiale.

Ce document et son contenu sont exclusifs à Illumina, Inc. et ses filiales (« Illumina »), et sont destinés à un usage contractuel de ses clients en lien avec l'utilisation du ou des produits décrits dans la présente et à aucune autre utilisation. Ce document et son contenu ne seront pas utilisés ou distribués dans tout autre but et/ou autrement communiqués, divulgués ou reproduits de quelque manière que ce soit sans l'autorisation préalable et écrite d'Illumina. Par le biais de ce document, Illumina ne fournit aucune licence sur ses droits de brevets, de marques, de copyrights ou tout autre droit de fait, ni n'en fournit sur de tels droits de tierces parties.

Les instructions présentes dans ce document doivent être strictement et explicitement respectées par le personnel qualifié et correctement formé afin d'assurer une utilisation correcte et sécuritaire du ou des produits décrits dans la présente. Tout le contenu de ce document doit être entièrement lu et compris avant d'utiliser le ou les produits.

LE DÉFAUT DE LIRE ET RESPECTER EXPLICITEMENT TOUTES LES INSTRUCTIONS CONTENUES DANS LA PRÉSENTE PEUT CAUSER DES DOMMAGES AU OU AUX PRODUITS, DES BLESSURES AUX PERSONNES, Y COMPRIS AUX UTILISATEURS OU AUTRES PERSONNES, ET DES DOMMAGES À D'AUTRES BIENS, ET ANNULERA TOUTE GARANTIE APPLICABLE AU OU AUX PRODUITS.

ILLUMINA N'EST PAS RESPONSABLE DE TOUT DOMMAGE CAUSÉ PAR UNE MAUVAISE UTILISATION DU OU DES PRODUITS DÉCRITS DANS LA PRÉSENTE (Y COMPRIS LES PARTIES DE CELLE-CI OU LE LOGICIEL).

© 2022 Illumina, Inc. Tous droits réservés.

Toutes les marques sont la propriété d'Illumina, Inc. ou de leurs propriétaires respectifs. Pour plus d'informations sur les marques, consultez [www.illumina.com/company/legal.html](http://www.illumina.com/company/legal.html).

# Table des matières

Historique des modifications .....	ii
Présentation .....	1
Méthodes d'analyse .....	1
Créer une analyse .....	5
Paramètres .....	7
Sorties d'analyse .....	8
Fichiers FASTQ .....	9
Fichiers BAM .....	9
Fichiers VCF .....	10
Afficher les résultats de l'analyse .....	16
<b>Assistance technique .....</b>	<b>17</b>

# Présentation

Le DRAGEN™ pour l'application Illumina® DNA Prep with Enrichment Dx effectue le démultiplexage, la génération de fichiers FASTQ, le mappage de lecture et l'alignement sur un génome de référence et la définition de variante en fonction du flux de travail d'analyse sélectionné.

## Méthodes d'analyse

DRAGEN for Illumina DNA Prep with Enrichment Dx effectue le démultiplexage, la génération FASTQ, la cartographie de lecture et l'alignement sur un génome de référence en fonction du flux de travail sélectionné :

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

## 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 l'analyse 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 l'analyse. 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 l'analyse.
- 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 un seul mésappariement par lecture d'index.

Le logiciel inclut la compression ORA pour compresser les fichiers FASTQ. 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

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. 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.

## Définition de variants germinaux DRAGEN

Le DRAGEN Germline Small Variant Caller prend les lectures d'ADN cartographiées et alignées comme entrée et définit les SNP et les indels grâce à une combinaison de détection par colonne et d'assemblage local *de novo* d'haplotypes.

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 autour de 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.

Dans le mode gVCF utilisé pour la définition de variants multi-échantillons évolutive, le DRAGEN Germline Small Variant Caller peut être exécuté par échantillon pour générer un fichier de définition de variant génomique intermédiaire (gVCF). Le gVCF peut ensuite être utilisé pour un génotypage conjoint efficace de plusieurs échantillons, ce qui permet le traitement incrémentiel rapide des échantillons et la mise à l'échelle à de grandes tailles de cohorte.

Étant donné que DRAGEN Germline Small Variant Caller dispose d'algorithmes qui lui permettent de distinguer efficacement les erreurs corrélées des vrais variants, les règles de filtrage sont très simples.

## Définition des variants somatiques

Le DRAGEN Somatic Small Variant Caller prend les lectures d'ADN cartographiées et alignées comme entrée et définit les SNV et les indels grâce à un assemblage local *de novo* d'haplotypes dans une région active.

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 autour de 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 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 la 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.



# Créer une analyse

Utilisez les étapes suivantes pour configurer une analyse dans Illumina Run Manager soit sur le NovaSeq 6000Dx ou à l'aide d'un navigateur sur un ordinateur en réseau. Les données d'échantillon peuvent être saisies manuellement ou en important une feuille d'échantillon.

## Paramètres de l'application et de l'analyse

1. Sur l'écran Runs (Analyses), sélectionnez **Create Run** (Créer une analyse).
2. Sélectionnez l'application DRAGEN for Illumina DNA Prep with Enrichment Dx, puis sélectionnez **Next** (Suivant).
3. Sur l'écran Run Settings (Paramètres de l'analyse), entrez un nom pour l'analyse. Le nom de l'analyse identifie l'analyse depuis le séquençage jusqu'à l'analyse.
4. **[En option]** Entrez une description d'analyse pour identifier davantage l'analyse.
5. Assurez-vous que le kit de préparation de librairie sélectionné est un kit de préparation de librairie Illumina DNA Prep with Enrichment Dx.
6. Sélectionnez la trousse d'adaptateur d'index souhaitée.
7. Entrez la longueur de lecture.  
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.
8. **[En option]** Entrez un ID de tube de librairie.
9. Sélectionnez **Next** (Suivant).

## Données de l'échantillon

Utilisez le tableau de l'écran Sample Data (Données de l'échantillon) pour saisir manuellement les informations sur l'échantillon. Vous pouvez également sélectionner **Import Samples** (Importer des échantillons) pour télécharger les informations sur les échantillons. Pour plus d'informations sur l'importation d'informations sur les échantillons, reportez-vous à la section [Importer des échantillons à la page 6](#) (Importer des échantillons).

### Saisir les échantillons manuellement

1. Saisissez un ID d'échantillon unique dans le champ Sample ID (ID d'échantillon).
2. Utilisez **Plate - Well Position** (Plaque - Position du puits) pour sélectionner la position du puits. Les champs Index i7, Index 1, Index i5 et Index 2 se remplissent automatiquement.
3. **[En option]** Entrez le nom de la librairie.
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.
5. Sélectionnez **Next** (Suivant).

## Importer des échantillons

Un modèle (\*.csv) est disponible en téléchargement sur l'écran Sample Data (Données de l'échantillon) lors de la planification d'une analyse dans Illumina Run Manager à l'aide d'un navigateur sur un ordinateur en réseau.

1. Sélectionnez **Download Template** (Télécharger le modèle) pour télécharger un fichier CSV vierge.
2. À partir du fichier CSV, entrez les informations sur l'échantillon et enregistrez le fichier.  
Le fichier CSV de feuille d'échantillon comprend les colonnes de données suivantes : ID de l'échantillon, plaque - position du puits, nom de la librairie **facultatif**.
3. Sélectionnez **Import Samples** (Importer des échantillons) et accédez à l'emplacement du fichier CSV.
4. Sélectionnez **Next** (Suivant).

## Paramètres de l'analyse

1. Sélectionnez le flux de travail d'analyse souhaité.
  - Génération des fichiers FASTQ
  - Génération de fichier FASTQ et VCF pour les variants germinaux pour un flux de travail germinale
  - Génération de fichier FASTQ et VCF pour les variants somatiques pour un flux de travail somatique
2. **[En option]** Si souhaité, cochez la case **Generate ORA compressed FASTQs** (Générer des FASTQ compressés ORA) pour autoriser la compression ORA pour les fichiers FASTQ.
3. **[Flux de travail de génération de fichiers VCF]** Utilisez le menu déroulant **Manifest File Selection** (Sélection du fichier de manifeste) pour sélectionner un fichier de manifeste.  
Un fichier manifeste est une entrée requise pour le DRAGEN for Illumina DNA Prep with Enrichment Dx. Le manifeste est un fichier BED délimité par des tabulations (\*.bed) qui spécifie les noms et les emplacements des régions de référence ciblées.
4. **[Flux de travail de génération de fichiers pour les variants somatiques FASTQ et VCF]** Utilisez le menu déroulant **Noise File Selection** (Sélection du fichier de bruit) pour sélectionner un fichier de bruit.  
Un fichier BED 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 à la page 7](#).
5. Sélectionnez **Next** (Suivant).

## Analyse Examiner

1. Sur l'écran Review (Examiner), passez en revue les informations saisies sur les écrans Run Settings (Paramètres de l'analyse), Sample Data (Données de l'échantillon) et Analysis Settings (Paramètre de l'analyse).

2. Sélectionnez **Save** (Enregistrer).

L'analyse est enregistrée dans l'onglet Planned (Planifié) de l'écran Runs (Analyses).

## Paramètres

Sélectionnez l'application sur l'écran Applications pour afficher les paramètres actuels et modifier les paramètres.

### Configuration

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

- **Trousses de préparation de librairie** : affiche la trousse de préparation de librairie par défaut pour l'application. Cet paramètre ne peut pas être modifié.
- **Trousses d'adaptateur d'index Adapter Kits** : affiche la trousse d'adaptateur d'index par défaut pour l'application. Cet paramètre ne peut pas être modifié.
- **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 l'analyse
- **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 l'analyse 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 l'analyse.

### Autorisations

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

### Filtrage du bruit

Le filtrage du bruit systématique est disponible lors de l'utilisation du flux de travail somatique. Le filtre peut être utilisé en mode Tumeur normale, mais est particulièrement utile pour les analyses Tumeur uniquement lorsqu'une tumeur normale correspondante n'est pas disponible.

Le bruit systématique BED doit être généré à partir d'échantillons normaux. Il est recommandé de créer des fichiers de bruit systématiques qui sont spécifiques à la préparation de la librairie, au système de séquençage et au panel. Pour la génération de fichiers de bruit, il est recommandé d'utiliser environ 50 échantillons normaux.

# Sorties d'analyse

DRAGEN for Illumina DNA Prep with Enrichment Dx enregistre les informations suivantes dans le dossier d'analyse. Seuls les flux de travail des variants germinaux et somatiques produisent un PDF.

- 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

Les fichiers de sortie d'analyse suivants sont générés par l'application. Les fichiers exacts générés dépendent du flux de travail d'analyse utilisé. Les fichiers de sortie trouvent dans le dossier Analyse.

Fichier de sortie	Description
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. Si la compression ORA est sélectionnée, le nom du fichier en tient compte.
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.

Fichier de sortie	Description
Rapport d'indicateurs de l'analyse (*.csv)	Contient des mesures de qualité sur l'analyse, y compris le rendement total 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
- Le signe plus (+)
- Les scores de qualité Phred est identifié dans un format codé ASCII + 33

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

```
@Instrument:IDAnalyse: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 l'analyse. En mode multinoeud, 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.

- 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.
- 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 : Le marqueur du nom de l'amplicon, qui enregistre l'identifiant de l'amplicon associé à la lecture des fichiers d'index BAM (\*.bam.bai) fournit un index du fichier BAM correspondant.

## Fichiers VCF

Le format de fichier 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 <code>dbSNP.txt</code> 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.

En-tête	Description
ALT	Allèles qui diffèrent de la lecture de référence. Par exemple, une délétion d'un T unique est représentée TT 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

En-tête	Description
FILTRE	<p>Si tous les filtres sont passés, PASS (Réussite) s'inscrit dans la colonne Filter (Filtre).</p> <p>Les entrées FILTER (Filtre) possibles du flux de travail germinale incluent :</p> <ul style="list-style-type: none"> <li>• <b>DRAGENSnpHardQUAL</b> : appliqué si le score QUAL du variant SNP n'atteint pas le seuil</li> <li>• <b>DRAGENIndelHardQUAL</b> : appliqué si le score QUAL du variant indel n'atteint pas le seuil</li> <li>• <b>LowDepth</b> : site filtré, car la profondeur de couverture n'atteint pas le seuil</li> <li>• <b>LowGQ</b> : site filtré, car la qualité du génotype n'atteint pas le seuil</li> <li>• <b>PloidyConflict</b> : la définition de génotype du paramètre de définition de variant n'est pas compatible avec la ploïdie chromosomique</li> <li>• <b>base_quality</b> : site filtré, car la qualité de base médiane des lectures alt à ce locus n'atteint pas le seuil</li> <li>• <b>filtered_reads</b> : site filtré, car une trop grande fraction de lectures a été filtrée</li> <li>• <b>fragment_length</b> : 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</li> <li>• <b>low_depth</b> : site filtré, car la profondeur de lecture est trop faible</li> <li>• <b>low_frac_info_reads</b> : site filtré, car la fraction de lectures informatives est inférieure au seuil</li> <li>• <b>normal_depth</b> : site filtré, car la profondeur de lecture normale de l'échantillon est trop faible</li> <li>• <b>long_indel</b> : site filtré, car la longueur de l'indel est trop longue</li> <li>• <b>mapping_quality</b> : site filtré, car la qualité de cartographie médiane des lectures alt à ce locus n'atteint pas le seuil</li> <li>• <b>multiallelic</b> : site filtré, car plus de deux allèles alt passent la LD de la tumeur</li> <li>• <b>non_homref_normal</b> : site filtré, car le génotype normal de l'échantillon n'est pas une référence homozygote</li> <li>• <b>no_reliable_supporting_read</b> : site filtré, car il n'existe aucune lecture somatique fiable.</li> <li>• <b>panel_of_normals</b> : Vu dans au moins un échantillon dans le panel des échantillons normaux vcf</li> <li>• <b>read_position</b> : site filtré, car la médiane des distances entre le début/la fin de la lecture et ce lieu est inférieure au seuil</li> <li>• <b>RMxNRepeatRegion</b> : site filtré, car tout ou partie de l'allèle variant est une répétition de la référence</li> <li>• <b>strand_artifact</b> : site filtré en raison d'une distorsion sévère du brin</li> <li>• <b>str_contraction</b> : site filtré en raison d'une erreur de PCR suspectée où l'allèle alt est une unité répétée de moins que la référence</li> <li>• <b>too_few_supporting_reads</b> : site filtré, car il y a trop peu de lectures de soutien dans l'échantillon de tumeur</li> <li>• <b>weak_evidence</b> : le score de variant somatique n'atteint pas le seuil</li> </ul>



En-tête	Description
FILTRE (suite)	<p>Les entrées FILTER (Filtre) possibles du flux de travail somatiques incluent :</p> <ul style="list-style-type: none"> <li>• <b>base_quality</b> : site filtré, car la qualité de base médiane des lectures alt à ce locus n'atteint pas le seuil</li> <li>• <b>filtered_reads</b> : site filtré, car une trop grande fraction de lectures a été filtrée</li> <li>• <b>fragment_length</b> : 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</li> <li>• <b>low_depth</b> : site filtré, car la profondeur de lecture est trop faible</li> <li>• <b>low_frac_info_reads</b> : site filtré, car la fraction de lectures informatives est inférieure au seuil</li> <li>• <b>normal_depth</b> : site filtré, car la profondeur de lecture normale de l'échantillon est trop faible</li> <li>• <b>long_indel</b> : site filtré, car la longueur de l'indel est trop longue</li> <li>• <b>mapping_quality</b> : site filtré, car la qualité de cartographie médiane des lectures alt à ce locus n'atteint pas le seuil</li> <li>• <b>multiallelic</b> : site filtré, car plus de deux allèles alt passent la LD de la tumeur</li> <li>• <b>non_homref_normal</b> : site filtré, car le génotype normal de l'échantillon n'est pas une référence homozygote</li> <li>• <b>no_reliable_supporting_read</b> : site filtré, car il n'existe aucune lecture somatique fiable.</li> <li>• <b>panel_of_normals</b> : Vu dans au moins un échantillon dans le panel des échantillons normaux vcf</li> <li>• <b>read_position</b> : site filtré, car la médiane des distances entre le début/la fin de la lecture et ce lieu est inférieure au seuil</li> <li>• <b>RMxNRepeatRegion</b> : site filtré, car tout ou partie de l'allèle variant est une répétition de la référence</li> <li>• <b>strand_artifact</b> : site filtré en raison d'une distorsion sévère du brin</li> <li>• <b>str_contraction</b> : site filtré en raison d'une erreur de PCR suspectée où l'allèle alt est une unité répétée de moins que la référence</li> <li>• <b>too_few_supporting_reads</b> : site filtré, car il y a trop peu de lectures de soutien dans l'échantillon de tumeur</li> <li>• <b>weak_evidence</b> : le score de variant somatique n'atteint pas le seuil</li> <li>• <b>systematic_noise</b> : site filtré sur la base de preuves de bruit systématique dans les échantillons normaux</li> </ul>

En-tête	Description
INFO	<p>Les entrées INFO possibles du flux de travail germinale incluent :</p> <ul style="list-style-type: none"> <li>• <b>AC</b> : comptage des allèles dans les géotypes pour chaque allèle ALT dans le même ordre que celui indiqué.</li> <li>• <b>AF</b> : fréquence des allèles pour chaque allèle ALT dans le même ordre que celui indiqué.</li> <li>• <b>AN</b> : nombre total d'allèles dans les géotypes appelés.</li> <li>• <b>DB</b> : adhésion dbSNP.</li> <li>• <b>FS</b> : valeur p à l'échelle de Phred utilisant le test exact de Fisher pour détecter la distorsion du brin.</li> <li>• <b>QD</b> : confiance/qualité par profondeur pour le variant.</li> <li>• <b>R2_5P_bias</b> : score basé sur la distorsion intersexuelle et la distance à partir de 5.</li> <li>• <b>SOR</b> : rapport de cotes symétrique du tableau de contingence 2x2 pour détecter la distorsion du brin.</li> <li>• <b>DP</b> : profondeur de lecture approximative (informative et non informative) ; certaines lectures peuvent avoir été filtrées en fonction de mapq, etc.</li> <li>• <b>END</b> : position d'arrêt de l'intervalle.</li> <li>• <b>FractionInformativeReads</b> : fraction de lectures informatives sur le nombre total de lectures.</li> <li>• <b>MQ</b> : qualité de mappage RMS.</li> <li>• <b>MQRankSum</b> : score Z du test de la somme des rangs de Wilcoxon des qualités de mappage de lecture Alt vs. Ref.</li> <li>• <b>ReadPosRankSum</b> : Z-score du test de somme des rangs de Wilcoxon du biais de position de lecture Alt vs. Ref.</li> <li>• <b>SOMATIC</b> : au moins un variant à cette position est un variant somatique.</li> </ul> <p>Les entrées INFO possibles du flux de travail somatiques incluent :</p> <ul style="list-style-type: none"> <li>• <b>DP</b> : profondeur de lecture approximative (informative et non informative) ; certaines lectures peuvent avoir été filtrées en fonction de mapq, etc.</li> <li>• <b>END</b> : position d'arrêt de l'intervalle.</li> <li>• <b>FractionInformativeReads</b> : fraction de lectures informatives sur le nombre total de lectures.</li> <li>• <b>MQ</b> : qualité de mappage RMS.</li> <li>• <b>MQRankSum</b> : score Z du test de la somme des rangs de Wilcoxon des qualités de mappage de lecture Alt vs. Ref.</li> <li>• <b>ReadPosRankSum</b> : Z-score du test de somme des rangs de Wilcoxon du biais de position de lecture Alt vs. Ref.</li> <li>• <b>AQ</b> : score de bruit systématique.</li> <li>• <b>hotspot</b> : site somatique connu, utilisé pour augmenter la confiance dans la définition.</li> <li>• <b>SOMATIC</b> : au moins un variant à cette position est un variant somatique.</li> </ul>

En-tête	Description
FORMAT	<p data-bbox="308 279 1383 352">La colonne Format dresse la liste des champs séparés par deux-points. Par exemple, GT:GQ.</p> <p data-bbox="308 363 1383 394">Les champs disponibles du flux de travail germinale incluent :</p> <ul data-bbox="308 405 1383 1392" style="list-style-type: none"> <li>• <b>AD</b> : profondeurs alléliques (en ne comptant que les lectures informatives sur le total des lectures) pour les allèles ref et alt dans l'ordre indiqué.</li> <li>• <b>AF</b> : fractions d'allèles pour les allèles alt dans l'ordre indiqué.</li> <li>• <b>DP</b> : profondeur de lecture approximative (les lectures avec MQ=255 ou de mauvais appariements sont filtrées).</li> <li>• <b>F1R2</b> : nombre de lectures dans l'orientation de la paire F1R2 prenant en charge chaque allèle.</li> <li>• <b>F2R1</b> : nombre de lectures dans l'orientation de la paire F2R1 prenant en charge chaque allèle.</li> <li>• <b>GP</b> : probabilités postérieures à l'échelle de Phred pour les génotypes tels que définis dans la spécification VCF.</li> <li>• <b>GQ</b> : qualité du génotype.</li> <li>• <b>GT</b> : 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.</li> <li>• <b>MB</b> : statistiques de composants par échantillon pour détecter le biais intersexuel.</li> <li>• <b>PL</b> : probabilités normalisées à l'échelle Phred pour les génotypes tels que définis dans la spécification VCF.</li> <li>• <b>PRI</b> : probabilités antérieures à l'échelle de Phred pour les génotypes.</li> <li>• <b>PS</b> : 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.</li> <li>• <b>SB</b> : statistiques de composants par échantillon qui comprennent le test exact de Fisher pour détecter la distorsion du brin.</li> <li>• <b>SQ</b> : qualité pour le variant somatique.</li> </ul> <p data-bbox="308 1402 1383 1434">Les champs disponibles du flux de travail somatique incluent :</p> <ul data-bbox="308 1444 1383 1894" style="list-style-type: none"> <li>• <b>AD</b> : profondeurs alléliques (en ne comptant que les lectures informatives sur le total des lectures) pour les allèles ref et alt dans l'ordre indiqué.</li> <li>• <b>AF</b> : fractions d'allèles pour les allèles alt dans l'ordre indiqué.</li> <li>• <b>DP</b> : profondeur de lecture approximative (les lectures avec MQ=255 ou de mauvais appariements sont filtrées).</li> <li>• <b>F1R2</b> : nombre de lectures dans l'orientation de la paire F1R2 prenant en charge chaque allèle.</li> <li>• <b>F2R1</b> : nombre de lectures dans l'orientation de la paire F2R1 prenant en charge chaque allèle.</li> <li>• <b>GT</b> : 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.</li> </ul>

En-tête	Description
FORMAT (suite)	<ul style="list-style-type: none"> <li>• <b>MB</b> : statistiques de composants par échantillon pour détecter le biais intersexuel.</li> <li>• <b>PS</b> : 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.</li> <li>• <b>SB</b> : statistiques de composants par échantillon qui comprennent le test exact de Fisher pour détecter la distorsion du brin.</li> <li>• <b>SQ</b> : qualité pour le variant somatique.</li> </ul>
SAMPLE	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.

## Afficher les résultats de l'analyse

Les analyses en cours sont affichées dans l'onglet Active (Actif). Les analyses terminées sont affichées dans l'onglet Completed (Terminé). Reportez-vous à [Documentation produit NovaSeq 6000Dx \(document n° 200010105\)](#) pour plus d'informations sur l'affichage des résultats.

# Assistance technique

Pour une assistance technique, contactez le support technique Illumina.

Site Web : [www.illumina.com](http://www.illumina.com)  
E-mail : [techsupport@illumina.com](mailto:techsupport@illumina.com)

## Numéros de téléphone du support technique Illumina

Région	Gratuit	International
Australie	+61 1800 775 688	
Autriche	+43 800 006249	+43 1 9286540
Belgique	+32 800 77 160	+32 3 400 29 73
Canada	+1 800 809 4566	
Chine		+86 400 066 5835
Danemark	+45 80 82 01 83	+45 89 87 11 56
Finlande	+358 800 918 363	+358 9 7479 0110
France	+33 8 05 10 21 93	+33 1 70 77 04 46
Allemagne	+49 800 101 4940	+49 89 3803 5677
Hong Kong, Chine	+852 800 960 230	
Inde	+91 8006500375	
Indonésie		0078036510048
Irlande	+353 1800 936608	+353 1 695 0506
Italie	+39 800 985513	+39 236003759
Japon	+81 0800 111 5011	
Malaisie	+60 1800 80 6789	
Pays-Bas	+31 800 022 2493	+31 20 713 2960
Nouvelle-Zélande	+64 800 451 650	
Norvège	+47 800 16 836	+47 21 93 96 93
Philippines	+63 180016510798	
Singapour	1 800 5792 745	
Corée du Sud	+82 80 234 5300	
Espagne	+34 800 300 143	+34 911 899 417

Région	Gratuit	International
Suède	+46 2 00883979	+46 8 50619671
Suisse	+41 800 200 442	+41 56 580 00 00
Taiïwan, Chine	+886 8 06651752	
Thaïlande	+66 1800 011 304	
Royaume-Uni	+44 800 012 6019	+44 20 7305 7197
États-Unis	+1 800 809 4566	+1 858 202 4566
Vietnam	+84 1206 5263	

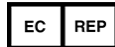
**Fiches de données de sécurité (FDS)** : disponibles sur le site Web d'Illumina à l'adresse [support.illumina.com/sds.html](https://support.illumina.com/sds.html).

**Documentation sur les produits** : disponible en téléchargement sur [support.illumina.com](https://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

CE



Illumina Netherlands B.V.  
Steenoven 19  
5626 DK Eindhoven  
Pays-Bas

**Commanditaire australien**

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

DESTINÉ AU DIAGNOSTIC IN VITRO UNIQUEMENT

© 2022 Illumina, Inc. Tous droits réservés.

**illumina**<sup>®</sup>